We Built a Fake News & Click-bait Filter: What Happened Next Will Blow Your Mind!

Georgi Karadzhov[1], Pepa Gencheva[1], Preslav Nakov[2], and Ivan Koychev[1]

[1]Sofia University "St. Kliment Ohridski", Bulgaria
[2]Qatar Computing Research Institute, HBKU, Qatar
{georgi.m.karadjov, pepa.k.gencheva}@gmail.com,
pnakov@hbku.edu.qa, koychev@uni-sofia.bg

Abstract

It is completely amazing! Fake news and click-baits have totally invaded the cyber space. Let us face it: everybody hates them for three simple reasons. Reason #2 will absolutely amaze you. What these can achieve at the time of election will completely blow your mind! Now, we all agree, this cannot go on, you know, somebody has to stop it. So, we did this research on fake news/click-bait detection and trust us, it is totally great research, it really is! Make no mistake. This is the best research ever! Seriously, come have a look, we have it all: neural networks, attention mechanism, sentiment lexicons, author profiling, you name it. Lexical features, semantic features, we absolutely have it all. And we have totally tested it, trust us! We have results, and numbers, really big numbers. The best numbers ever! Oh, and analysis, absolutely top notch analysis. Interested? Come read the shocking truth about fake news and click-bait in the Bulgarian cyber space. You won't believe what we have found!

1 Introduction

Fake news are written and published with the intent to mislead in order to gain financially or politically, often targeting specific user groups. Another type of harmful content on the Internet are the so-called *click-baits*, which are distinguished by their sensational, exaggerated, or deliberately false headlines that grab attention and deceive the user into clicking an article with questionable content.

While the motives behind these two types of fake news are different, they constitute a growing problem as they constitute a sizable fraction of the online news that users encounter on a daily basis. With the recent boom of Internet, mobile, and social networks, the spread of fake news increases exponentially. Using on-line methods for spreading harmful content makes the task of keeping the Internet clean significantly harder as it is very easy to publish an article and there is no easy way to verify its veracity. Currently, domains that consistently spread misinformation are being banned from various platforms, but this is a rather inefficient way to deal with fake news as websites that specialize in spreading misinformation are reappearing with different domain names. That is why our method is based purely on text analysis,[1] without taking into account the domain name or website's reliability as a source of information. Our work is focused on exploring various stylistic and lexical features in order to detect misleading content, and on experiments with neural network architectures in order to evaluate how deep learning can be used for detecting fake news. Moreover, we created various language-specific resources that could be used in future work on fake news and clickbait detection for Bulgarian, including task-specific word embeddings and various lexicons and dictionaries extracted from the training data.[2]

[1]An earlier version of the system participated in the *Hack the fake news* hackathon, where it was ranked best in terms of classification accuracy and robustness. See the official results here: https://gitlab.com/datasciencesociety/case_fake_news/blob/master/Teams_Final_Score.xlsx

[2]The implementation of the final system that we present in this paper is available at https://github.com/lachezarbozhkov/hack_the_fake_news

Proceedings of Recent Advances in Natural Language Processing, pages 334–343,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_045

2 Related Work

Trustworthiness and veracity analytics of online statements is an emerging research direction (Rowe and Butters, 2009). This includes predicting credibility of information shared in social media (Mitra et al., 2017), stance classification (Zubiaga et al., 2016a) and contradiction detection in rumours (Lendvai and Reichel, 2016). For example, Castillo et al. (2011) studied the problem of finding false information about a newsworthy event. They compiled their own dataset, focusing on tweets using a variety of features including user reputation, author writing style, and various time-based features. Canini et al. (2011) analysed the interaction of content and social network structure, and Morris et al. (2012) studied how Twitter users judge truthfulness. They found that this is hard to do based on content alone, and instead users are influenced by heuristics such as user name.

Rumour detection in social media represents yet another angle of information credibility. Zubiaga et al. (2015) studied how people handle rumours in social media. They found that users with higher reputation are more trusted, and thus can spread rumours among other users without raising suspicions about the credibility of the news or of its source. Lukasik et al. (2015) and Ma et al. (2015) used temporal patterns to detect rumours and to predict their frequency, Zubiaga et al. (2016b) focused on conversational threads, and Karadzhov et al. (2017) used deep learning to verify claims using the Web as a corpus.

Veracity of information has been also studied in the context of online personal blogs (Johnson et al., 2007), community question answering forums (Nakov et al., 2017), and political debates (Gencheva et al., 2017).

Astroturfing and misinformation detection represent another relevant research direction. Their importance is motivated by the strong interest from political science, and research methods are driven by the presence of massive streams of micro-blogging data, e.g., on Twitter (Ratkiewicz et al., 2011). While astroturfing has been primarily studied in microblogs such as Twitter, here we focus on on-line news and click-baits instead.

Identification of malicious accounts in social networks is another related research direction. This includes detecting *spam accounts* (Almaatouq et al., 2016; Mccord and Chuah, 2011), *fake accounts* (Fire et al., 2014; Cresci et al., 2015), *compromised accounts* and *phishing accounts* (Adewole et al., 2017). *Fake profile detection* has also been studied in the context of cyber-bullying (Galán-García et al., 2014). A related problem is that of *Web spam detection*, which was addressed as a text classification problem (Sebastiani, 2002), e.g., using spam keyword spotting (Dave et al., 2003), lexical affinity of arbitrary words to spam content (Hu and Liu, 2004), frequency of punctuation and word co-occurrence (Li et al., 2006).

Fake news detection is most closely related to the present work. While social media have been seen for years as the main vehicle for spreading information of questionable veracity, recently there has been a proliferation of fake news, often spread on social media, but also published in specialized websites. This has attracted research attention recently. For example, there has been work on studying credibility, trust, and expertise in news communities (Mukherjee and Weikum, 2015). The credibility of the information published in online news portals has been questioned by a number of researchers (Brill, 2001; Ketterer, 1998; Finberg et al., 2002). As timing is crucial when it comes to publishing breaking news, it is simply not possible to double-check the facts and the sources, as is usually standard in respectable printed newspapers and magazines. This is one of the biggest concerns about on-line news media that journalists have (Cassidy, 2007). Finally, Conroy et al. (2015) review various methods for detecting fake news, e.g., using linguistic analysis, discourse, linked data, and social network features.

All the above work was for English. The only work on fact checking for Bulgarian is that of (Hardalov et al., 2016), but they focused on distinguishing serious news from humorous ones. In contrast, here we are interested in finding news that are not designed to sound funny, but to make the reader believe they are real. Unlike them, we use a deep learning approach.

3 Fake News & Click-bait Dataset

We use a corpus of Bulgarian news over a fixed period of time, whose factuality had been questioned. The news come from 377 different sources from various domains, including politics, interesting facts and tips&tricks. The dataset was prepared for the *Hack the Fake News* hackathon. It was provided by the Bulgarian Association of PR Agencies[3] and is available in Gitlab[4]. The corpus was automatically collected, and then annotated by students of journalism. Each entry in the dataset consists of the following elements: URL of the original article, date of publication, article heading, article content, a label indicating whether the article is fake or not, and another label indicating whether it is a click-bait.

The training dataset contains 2,815 examples, where 1,940 (i.e., 69%) are fake news and 1,968 (i.e., 70%) are click-baits; we further have 761 testing examples. However, there is 98% correlation between fake news and click-baits, i.e., a model trained on fake news would do well on click-baits and vice versa. Thus, below we focus on fake news detection only.

One important aspect about the training dataset is that it contains many repetitions. This should not be surprising as it attempts to represent a natural distribution of factual vs. fake news on-line over a period of time. As publishers of fake news often have a group of websites that feature the same deceiving content, we should expect some repetition.

In particular, the training dataset contains 434 unique articles with duplicates. These articles have three reposts each on average, with the most reposted article appearing 45 times. If we take into account the labels of the reposted articles, we can see that if an article is reposted, it is more likely to be fake news. The number of fake news that have a duplicate in the training dataset are 1018 whereas, the number of articles with genuine content that have a duplicate article in the training set is 322. We detect the duplicates based on their titles as far as they are distinctive enough and the content is sometimes slightly modified when reposted.

This supports the hypothesis that fake news websites are likely to repost their content. This is also in line with previous research (Ma et al., 2015), which has found it beneficial to find a pattern of how a rumour is reposted over time.

4 Method

We propose a general framework for finding fake news focusing on the text only. We first create some resources, e.g., dictionaries of words strongly correlated with fake news, which are needed for feature extraction. Then, we design features that model a number of interesting aspects about an article, e.g., style, intent, etc. Moreover, we use a deep neural network to learn task-specific representations of the articles, which includes an attention mechanism that can focus on the most discriminative sentences and words.

4.1 Language Resources

As our work is the first attempt at predicting click-baits in Bulgarian, it is organized around building new language-specific resources[5] and analyzing the task.

Word embeddings: We train 300-dimensional domain-specific word embeddings using word2vec (Mikolov et al., 2013) on 100,000 Bulgarian news articles from the same sources as the main dataset. The labelled dataset we use in our system is a subset of these articles. Finally, we end up with 207,270 unique words that occur in five or more documents. We use these embeddings for text representation, and as an input to our attention-based nevural network.

Latent Dirichlet allocation (LDA): We use LDA (Blei et al., 2003) in order to build domain-specific topic models, which could be useful for inducing classes of words that signal fake/factual news. The LDA model is trained on the same 100,000 Bulgarian news articles as for training the word embeddings. In our experiments, these LDA classes proved helpful by themselves, but they did not have much to offer on top of the word embeddings. Thus, we ended up not using them in our final system, but we chose to still release them as other researchers might find them useful in the future.

[3] http://www.bapra.bg/

[4] https://gitlab.com/datasciencesociety/case_fake_news/tree/master/data

[5] We make these resources freely available in order to promote reproducibility and to enable future research: https://github.com/gkaradzhov/ClickbaitRANLP

Fact-checking lexicon: Using lexicons of sentiment words has been shown to be very successful for the task of sentiment analysis (Mohammad and Turney, 2013), and we applied the same idea to extract a *fact-checking lexicon*. In particular, we use point-wise mutual information (PMI) to find terms (words, word bi-grams, and named entities) that are highly correlated with the fake/factual news class.

We calculated the PMI scores for uni-grams, bi-grams and on extracted named entities. Table 1 shows some of the most significant words for the fake news class. We can see in the table some words that grab people attention, but are not very informative by themselves, such as *mysterious* or *phenomenon*. These words are largely context-independent and are likely to remain stable in their usage across different domains and even over an extended period of time. Thus, they should be useful beyond this task and this dataset.

Other lexicons: Finally, we create four lexicons that can help to model the difference in language use between fake and factual news articles. In particular, we explored and merged/cleansed a number of on-line resources in order to put together the following lexicons: (*i*) common typos in Bulgarian written text, (*ii*) Bulgarian slang words, (*iii*) commonly used foreign words, and (*iv*) English words with Bulgarian equivalents. We separate the latter two, because of the frequent usage of English words in common language. We make these lexicons freely available for future research.

4.2 Features

4.2.1 Stylometric Features

Fake news are written with the intent to deceive, and their authors often use a different style of writing compared to authors that create genuine content. This could be either deliberately, e.g., if the author wants to adapt the text to a specific target group or wants to provoke some particular emotional reaction in the reader, or unintentionally, e.g., because the authors of fake news have different writing style and personality compared to journalists in mainstream media. Disregarding the actual reason, we use features from author profiling and style detection (Rangel et al., 2013).

Original word	Translation	PMI
chemtrails	chemtrails	0.92
феноменните	the phenomenal	0.94
следете в	follow in	0.97
тайнствена	mysterious	0.95
скрит	hidden	0.84

Table 1: Words most strongly associated with the fake news class.

Use of specific words that have strong correlation with one of the classes (48 features). We used the above-described PMI-based fact-checking lexicons to extract features based on the presence of lexicon words in the target article. We end up with the following features: 16 for uni-grams + 16 for bi-grams + 16 for named entities, where we have a feature for the sum and also for the average of the word scores for each of the target classes (click-bait, non-click-bait, fake, non-fake), and we had these features separately for the title and for the body of the article.

Readability index (4 features): We calculate standard readability metrics including the type-token ratio, average word length, Flesch–Kincaid readability test (Kincaid et al., 1975) and Gunning-Fog index (Gunning, 1952). The last two metrics give scores to the text corresponding to the school grade the reader of the target article should have in order to be able to read and understand it easily. These metrics use statistics about the number of syllables, the number of words, and their length.

Orthographic features (12 features): The orthographic features used in our system include: the number of words in the title and in the content; the number of characters in the title and in the content; the number of specific symbols in the title and in the content, counting the following as symbols $.!;#?:-+%&̂(), ; the number of capital letters in the title and in the content; the fraction of capital letters to all letters in the title and in the content; the number of URLs in the content; the overlap between the words from the title and the words of the content, relying on the fact that click-baits tend to have content that does not quite match their title. These features can be very effective for modelling the author's style.

Use of irregular vocabulary (4 features): During the initial analysis of our training dataset, we noticed the presence of a high number of foreign words. As it is not common in Bulgarian news articles to use words in another language, we thought that their presence could be a valuable feature to use. One of the reasons for their occurrence might be that they were translated from a foreign resource, or that they were borrowed. We further found that many articles that were labelled as fake news contained a high number of slang words, and we added this as a feature as well. Finally, we have a feature that counts the typos in the text.

4.2.2 Lexical Features

General lexical features are often used in natural language processing as they are somewhat task-independent and reasonably effective in terms of classification accuracy. In our experiments, we used TF.IDF-based features over the title and over the content of the article we wanted to classify. We had these features twice – once for the title and once for the the content of the article, as we wanted to have two different representations of the same article. Thus, we used a total of 1,100 TF.IDF-weighted features (800 content + 300 title), limiting the vocabulary to the top 800 and 300 words, respectively (which occurred in more than five articles). We should note that TF.IDF features should be used with caution as they may not remain relevant over time or in different contexts without retraining.

4.2.3 Grammatical Features

The last type of hand-crafted features that we used are the grammatical features. First, we evaluate how often stop words are used in the content of the article. Extensive usage of stop words may indicate irregularities in the text, which would be missed by the above features. Additionally, we extract ten coarse-grained part-of-speech tags from the content of the article and we use part-of-speech occurrence ratios as features. This makes a total of twenty features, as we have separate features for the title and for the contents.

4.2.4 Semantic Features

All the above features are hand-crafted, evaluating a specific text metric or checking whether specific words highly correlate with one of the classes. However, we lack features that target the semantic representation of the text itself. Thus, we further use two types of word representations.

Word embeddings (601 features). As we said above, we trained domain-specific word embeddings. In order to incorporate them as features, we calculate the average vector for the title and separately for the content of the news article. We end up with two 300-dimensional embedding representations of the semantics of the articles, which we use as 300+300=600 features. We also compute the cosine similarity between the average vector of the title and the average vector of the content, because we believe that this is a highly indicative measure for at least click-bait articles, whose content differs from what their title says.

Task-specific embeddings. As a more advanced representation, we feed the text into an attention-based deep neural network, which we train to produce a task-specific embedding of the news articles. The network is designed to recognize words and sentences that contribute to the click-bait class attribution. The architecture is described in details in Section 4.4.1

4.3 Some Features that we Ignored

As we mentioned above, our method is purely text-based. Thus, we ignored the publishing date of the article. In future work, it could be explored as a useful piece of information about the credibility of the article, as there is interesting research in this direction (Ma et al., 2015). We also disregarded the article source (the URL) because websites that specialize in producing and distributing fake content are often banned and then later reappear under another name. We recognize that the credibility of a specific website could be a very informative feature, but, for the sake of creating a robust method for fake news detection, our system relies only on the text when predicting whether the target article is likely to be fake. We describe our features in more detail below.

4.4 Model

Our framework for fake news detection is comprised of two components, which are used one after the other. First, we have an attention-based deep neural network model, which focuses on the segments of the text that are most indicative of the target class identification, and as a side effect learns task-specific representations of the news articles. We extract these representations from the last hidden layer in the network, and we feed it to the SVM classifier together with the hand-crafted features.

4.4.1 Attention Mechanism

The attention network (Hermann et al., 2015), (Yang et al., 2016) is a powerful mechanism, inspired by the human ability to spot important sections in images or text. We adopt the approach used in (Rocktäschel et al., 2015) and employ an attention neural networks to build attention over the text of a piece of news with respect to the title it has. As far as it is in the nature of click-baits to have titles that are different from the text of the news, the attentional layers of the neural network should spot when the two texts talk about the same thing and when they are not corresponding or accurate. We implemented the attention mechanism using Keras (Chollet et al., 2015) with the Tensorflow back-end (Abadi et al., 2015).

The architecture of the network with attention layers is shown in Figure 1. Our neural model is based on Gated Recurrent Units (GRUs). GRUs are gating mechanism in RNNs which provide the ability to learn long-term dependencies and were first introduced in (Cho et al., 2014). Given the document embedding, the GRUs build representations using input and forget gates, which help storing the valuable information through time. They build embeddings of the title and the text of the news, where at each step the unit has information only about the output from the previous step. This can be considered as a drawback, as far as we would considerably benefit if each step could construct its decision based not only on the previous step's output, but on all of the words that were processed so far. To improve this, the attention layer, for each step in the text sequence, uses the output of the steps in the title sequence. Thus, the layer learns weights, designating the strength of the relatedness between each word in the title and each word in the content.

For the neural network, we take the first 50 symbols of the title and the content of the news, which we choose after experiments. We train the neural network for 20 epochs and the final classification is derived with sigmoid activation. The optimizer used for the training is Adam optimizer. We feed the neural network with the embedding of the words we built earlier with word2vec.

As we will see below, the neural network is inferior in terms of performance to a feature-rich SVM (even though it performs well above the baseline). This is because it only has access to word embeddings, and does not use the manually-crafted features. Yet, its hidden layer represents a 128-dimensional task-specific embedding of the input article, and it turns out that using it as a list of 128 features in the SVM classifier yields even further great improvement, as we will see below. In this way, we combine a deep neural network with an attention mechanism with kernel-based SVM.

Features	P	R	F1	Acc
Lexical	75.53	74.59	75.02	79.89
Stylometric	74.35	65.99	67.68	77.52
Grammatical	73.23	50.60	42.99	71.48
Embeddings	61.48	53.95	51.67	71.22

Table 2: Performance of the individual groups of hand-crafted features.

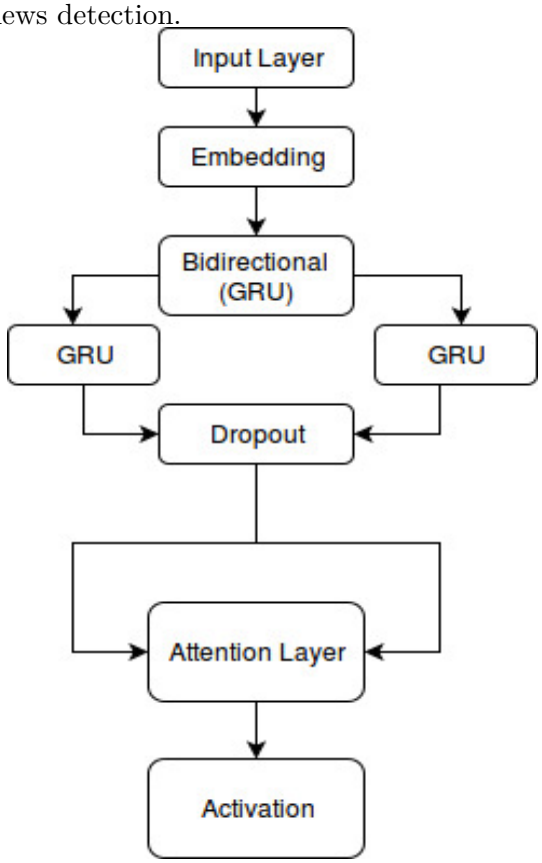
4.4.2 SVM

We feed the above-described hand-crafted features together with the task-specific embeddings learned by the deep neural neural network (a total of 1,892 attributes combined) into a Support Vector Machines (SVM) classifier (Cortes and Vapnik, 1995). SVMs have proven to perform well in different classification settings, including in the case of small and noisy datasets.

5 Experiments and Evaluation

We trained on the 2,815 training examples, and we tested on the 761 testing ones. The test dataset was provided apart from the training one, thus we didn't have to partition the

Figure 1: The architecture of our hierarchical attention deep neural network for click-bait news detection.



original dataset to receive a testing one. The validation of the models was performed on a randomly chosen subset of sentences - one fifth of the original set. We scaled each feature individually by its maximum absolute value to end up with each feature having values in the [0;1] interval. We used an RBF kernel for the SVM, and we tuned the values of C and γ using cross-validation. We trained the neural network using RMSProp (Tieleman and Hinton, 2012) with a learning rate of 0.001 and mini-batches of size 32, chosen by performing experiments with cross-validation . We evaluated the model after each epoch and we kept the one that performed best on the development dataset.

Table 2 shows the performance of the features in groups as described in Section 4.2. We can see that, among the hand-crafted features, the lexical features yield the best results, i.e., words are the most indicative features. The good results of the stylometric features indicate that the intricacies of language use are highly discriminative. The next group is the one with the grammatical features, which shows good performance in terms of Precision. The last one are the embedding features, which although having low individual performance, contribute to the overall performance of the system as shown in next paragraph.

Evaluating the final model, we set as a baseline the prediction of the majority class, i.e., the fake news class. This baseline has an F1 of 41.59% and accuracy of 71.22%. The performance of the built models can be seen in Table 3. Another stable baseline, apart from just taking the majority class, is the TF.IDF bag-of-words approach, which sets a high bar for the general model score. We then observe how much the attention mechanism embeddings improve the score (AtNN). Finally, we add the hand-crafted features (Feats), which further improve the performance. From the results, we can conclude that both the attention-based task-specific embeddings and the manual features are important for the task of finding fake news.

6 Conclusion and Future Work

We have presented the first attempt to solve the fake news problem for Bulgarian. Our method is purely text-based, and ignores the publication date and the source of the article. It combines task-specific embeddings, produced by a two-level attention-based deep neural network model, with manually crafted features (stylometric, lexical, grammatical, and semantic), into a kernel-based SVM classifier. We further produced and shared a number of relevant language resources for Bulgarian, which we created for solving the task.

The evaluation results are encouraging and suggest the potential applicability of our approach in a real-world scenario. They further show the potential of combining attention-based task-specific embeddings with manually crafted features. An important advantage of the attention-based neural networks is that the produced representations can be easily visualized and potentially interpreted as shown in (Hermann et al., 2015). We consider the implementation of such visualization as an important future work on the task.

Feature Group	P	R	F1	Acc
Baseline	35.61	50.00	41.59	71.22
TF.IDF	75.53	74.59	75.02	79.89
AttNN	78.52	78.74	78.63	81.99
TF.IDF &AttNN	79,89	79.40	79.63	83.44
TF.IDF &Feats &AttNN	80.07	79.49	79.77	83.57

Table 3: Performance of different models: AttNN – Attention Neural Network, Feats – handcrafted features.

Acknowledgements

We would like to thank Lachezar Bozhkov, who was part of our team in the *Hack the Fake News* hackathon, for his insight. This work is supported by the NSF of Bulgaria under Grant No. DN-02/11/2016 - ITDGate.

References

Martín Abadi, Ashish Agarwal, Paul Barham, and et. al. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. http://tensorflow.org/.

Kayode Sakariyah Adewole, Nor Badrul Anuar, Amirrudin Kamsin, Kasturi Dewi Varathan, and Syed Abdul Razak. 2017. Malicious accounts: Dark of the social networks. *Journal of Network and Computer Applications* 79:41–67.

Abdullah Almaatouq, Erez Shmueli, Mariam Nouh, Ahmad Alabdulkareem, Vivek K Singh, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, et al. 2016. If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. *International Journal of Information Security* 15(5):475–491.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Ann M Brill. 2001. Online journalists embrace new marketing function. *Newspaper Research Journal* 22(2):28.

K. R. Canini, B. Suh, and P. L. Pirolli. 2011. Finding credible information sources in social networks based on content and social structure. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. pages 1–8.

William P Cassidy. 2007. Online news credibility: An examination of the perceptions of newspaper journalists. *Journal of Computer-Mediated Communication* 12(2):478–498.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*. Hyderabad, India, WWW '11, pages 675–684.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* .

François Chollet et al. 2015. Keras. `https://github.com/fchollet/keras`.

Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake .

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20(3):273–297. https://doi.org/10.1023/A:1022627411411.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2015. Fame for sale: efficient detection of fake twitter followers. *Decision Support Systems* 80:56–71.

Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Web conference*. Budapest, Hungary, WWW '03, pages 519–528.

Howard Finberg, Martha L Stone, and Diane Lynch. 2002. Digital journalism credibility study. *Online News Association. Retrieved November* 3:2003.

Michael Fire, Dima Kagan, Aviad Elyashar, and Yuval Elovici. 2014. Friend or foe? fake profile identification in online social networks. *Social Network Analysis and Mining* 4(1):1–23.

Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. 2014. Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying. In *Proceedings of*

the International Joint Conference SOCO'13-CISIS'13-ICEUTE'13, Springer International Publishing, Advances in Intelligent Systems and Computing, pages 419–428.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cede no, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the 2017 International Conference on Recent Advances in Natural Language Processing*. Varna, Bulgaria, RANLP '17.

Robert Gunning. 1952. The technique of clear writing .

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2016. In search of credible news. In *Proceedings of the 17th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Varna, Bulgaria, AIMSA '16, pages 172–180.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *CoRR* abs/1506.03340. http://arxiv.org/abs/1506.03340.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, Washington, USA, KDD '04, pages 168–177.

Thomas J Johnson, Barbara K Kaye, Shannon L Bichard, and W Joann Wong. 2007. Every blog has its day: Politically-interested internet users' perceptions of blog credibility. *Journal of Computer-Mediated Communication* 13(1):100–122.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cede no, and Ivan Koychev. 2017. Fully automated fact checking using external sources. In *Proceedings of the 2017 International Conference on Recent Advances in Natural Language Processing*. Varna, Bulgaria, RANLP '17.

Stan Ketterer. 1998. Teaching students how to evaluate and use online resources. *Journalism & Mass Communication Educator* 52(4):4.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Piroska Lendvai and Uwe D Reichel. 2016. Contradiction detection for rumorous claims. *arXiv preprint arXiv:1611.02588* .

Wenbin Li, Ning Zhong, and Chunnian Liu. 2006. Combining multiple email filters based on multivariate statistical analysis. In *Foundations of Intelligent Systems*, Springer, pages 729–738.

Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Point process modelling of rumour dynamics in social media. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 518–523.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. Melbourne, Australia, CIKM '15, pages 1751–1754.

Michael Mccord and M Chuah. 2011. Spam detection on twitter using traditional classifiers. In *International Conference on Autonomic and Trusted Computing*. Springer, pages 175–186.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Tanushree Mitra, G Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. In *Proc. CSCW*.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29(3):436–465.

Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM, New York, NY, USA, CSCW '12, pages 441–450.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, pages 353–362.

Preslav Nakov, Tsvetomila Mihaylova, Lluís Màrquez, Yashkumar Shiroya, and Ivan Koychev. 2017. Do not trust the trolls: Predicting credibility in community question answering forums. In *Proceedings of the 2017 International Conference on Recent Advances in Natural Language Processing*. Varna, Bulgaria, RANLP '17.

Francisco Rangel, Paolo Rosso, Moshe Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*. CELCT, pages 352–365.

Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*. Hyderabad, India, WWW '11, pages 249–252.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664* .

Matthew Rowe and Jonathan Butters. 2009. Assessing Trust: Contextual Accountability. In *Proceedings of the First Workshop on Trust and Privacy on the Social and Semantic Web*. Heraklion, Greece, SPOT '09.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)* 34(1):1–47.

Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4(2):26–31.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*. pages 1480–1489.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, Rob Procter, and Peter Tolmie. 2015. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *arXiv preprint arXiv:1511.07487* .

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016a. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *arXiv preprint arXiv:1609.09028* .

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016b. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE* 11(3):1–29.