Stance Detection in Fake News: A Combined Feature Representation

Bilal Ghanem
PRHLT Research Center,
Universitat Politècnica de València,
Spain
bigha@doctor.upv.es

Paolo Rosso
PRHLT Research Center,
Universitat Politècnica de València,
Spain
prosso@dsic.upv.es

Francisco Rangel
PRHLT Research Center,
Universitat Politècnica de València,
Autoritas Consulting,
Spain
francisco.rangel@autoritas.es

Abstract

With the uncontrolled increasing of fake news and rumors over the Web, different approaches have been proposed to address the problem. In this paper, we present an approach that combines lexical, word embeddings and n-gram features to detect the stance in fake news. Our approach has been tested on the Fake News Challenge (FNC-1) dataset. Given a news title-article pair, the FNC-1 task aims at determining the relevance of the article and the title. Our proposed approach has achieved an accurate result (59.6 % Macro F1) that is close to the state-of-the-art result with 0.013 difference using a simple feature representation. Furthermore, we have investigated the importance of different lexicons in the detection of the classification labels.

1 Introduction

Recently, many phenomena appeared and spread in the Internet, especially with the huge propagation of information and the growth of social networks. Some of these phenomena are fake news, rumors and misinformation. In general, the detection of these phenomena is crucial since in many situations they expose the people to danger[1]. Journalism made several efforts in addressing these problems by presenting a validity proof to the audience. Unfortunately, these manual attempts take much time and effort from the journalists and, at the same time, they cannot cover the rapid spread of these fake news. Hence, there is the need for addressing the problem from an automatic perspective. Fake news gained large attention recently from the natural language processing (NLP) research community and many approaches have been proposed. These approaches investigated fake news from network and textual perspectives (Shu et al., 2017). Some of the textual approaches handled the phenomenon from a validity aspect, where they labeled a claim as ”False”, ”True”, or ”Half-True”. Others tried to tackle it from a stance perspective, similar to stance detection works on Twitter (Mohammad et al., 2016; Taulé et al., 2017; Lai et al., 2018) that tried to determine whether a tweet is in favor, against, or neither to a given target entity (person, organization, etc.). Where in fake news, they replaced the tuple of the tweet and the target entity with a claim and an article; also a different stances' set is used (agree, disagree, discuss, and unrelated).

Several shared tasks have been proposed: Fake News Challenge (FNC-1) (Rao and Pomerleau, 2017), RumorEval (Derczynski et al., 2017), CheckThat (Barrón-Cedeño et al., 2018), and Fact Extraction and Verification (FEVER)[2]. In FNC-1, the organizers proposed the task to be approached from a stance perspective; the goal is to predict how other articles orient to a specific fact, similarly than in RumorEval (task-A). While in both RumorEval (task-B) and CheckThat (task-B) a rumor/claim has been submitted and the task objective is to validate the truthfulness of this sentence (true, half-true, or false). In the first task of CheckThat (task-A) participants were asked to detect claims that are worthy for checking (may have facts), as preliminary step to task B. Finally, the purpose of FEVER shared task is to evaluate the ability of a system to verify a factual claim using evidences from Wikipedia, where each re-

[1] https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate

[2] http://fever.ai/task.html

Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 66–71
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics

trieved evidence (in case there are many) should be labeled as ”Supported”, ”Refuted” or ”NotEnoughInfo” (if there isn’t sufficient evidence to either support or refute it). The given attention to fake news and rumors detection in the literature is more than the one gained by detecting worthy claims. The orientation in these works was towards inferring these worthy claims using linguistic and stylistic aspects (Ghanem et al., 2018c; Hassan et al., 2015).

2 Related Work

From an NLP perspective, many approaches proposed to employ statistical (Magdy and Wanas, 2010), linguistic (Markowitz and Hancock, 2014; Volkova et al., 2017), and stylistic (Potthast et al., 2017) features. Other approaches incorporated different combination of features, such as word or character n-grams overlapping score, bag-of-words (BOW), word embeddings, and latent semantic analysis features (Riedel et al., 2017; Hanselowski et al., 2017; Karadzhov et al., 2018). In some cases, authors used external features and retrieved evidences from the Web. For example, in (Ghanem et al., 2018b) the authors utilized both Google and Bing search engines to investigate the factuality of political claims. In (Mihaylov et al., 2015), a similar work has also retrieved evidences from Google and online blogs to validate sentences in question answering forums. In other attempts, some approaches utilized deep learning architectures to validate fake news. In (Baird et al., 2017), an approach combined a Convolutional Neural Network with a Gradient Boost classifier to predict the stance on FNC. As a result, their approach achieved the highest accuracy in the task results. Using a different deep learning architecture, the authors in (Hanselowski et al., 2018) used a Long Short-Term Memory (LSTM) network combined with other features such as bag-of-characters (BOC), BOW and topic model features based on non-negative matrix factorization, Latent Dirichlet Allocation, and Latent Semantic Indexing. They achieved state-of-the-art results (60.9% Macro F1) on the FNC-1 dataset.

The approaches that were proposed in both fake news and rumors detection are slightly different, since both phenomena were studied in different environment. Fake news datasets generally were collected from formal sources (political debates or Web news articles). On the other hand, Twitter was the source for rumors datasets. Therefore, the proposed approaches for rumors focused more on the propagation of tweets (ex. retweet ratio (Enayet and El-Beltagy, 2017)) and the writing style of the tweets (Kochkina et al., 2017).

3 Stance Detection in FNC-1

3.1 Task

Given a pair of text fragments (title and article) obtained from news, the task goal is to estimate the relative perspective (stance) of these two fragments with respect to a specific topic. In other words, the stance prediction of an article towards the title of this article. For each input pair, there are 4 stance labels: Agree, Disagree, Discuss, and Unrelated. ”Agree” if the article supports the title; ”disagree” if refuses it; ”discuss” whether the article discusses the title but without showing an in favor or against stance; and ”unrelated” when the article describes a different topic than the one of the title. The task’s dataset is imbalanced in a high ratio (see next section). Therefore, the organizers introduced a weighted accuracy score for the evaluation. Their proposed score gave 25% of the final score for predicting the unrelated class, while 75% for the other classes. Later, the authors in (Hanselowski et al., 2018) proposed an in-depth analysis to discuss FNC-1 experimental setup. They showed that this accuracy metric is not appropriate and fails to take into account the imbalanced class distribution, where models performing well on the majority class and poorly on the minority classes are favored. Therefore, they proposed Macro F1 metric to be used in this task. Accordingly, in this paper we show the experimental results using the Macro F1 measure.

3.2 Corpus

The presented dataset was built using 300 different topics. The training part consists of 49,972 tuples in a form of title, article, and label, while the test part consists of 25,413 tuples. The ratio of each label (class) in the dataset is: 73.13% Unrelated, 17.82% Discuss, 7.36% Agree, and 1.68% Disagree. Clearly the dataset is heavily biased towards the unrelated label. Titles length ranges between 8 and 40 words, whereas for the articles ranges between 600 and 7000 words (Bhatt et al., 2018). These numbers show a real challenge to predict the stance between these two fragments that are totally different in lengths.

3.3 Tough-to-beat Baseline

The organizers presented a tough baseline using Gradient Boost decision tree classifier. In contrast to other shared tasks, their baseline employed more sophisticated features. As features, they employed n-gram co-occurrence between the titles and articles using both character and word grams (using a combination of multiple lengths) along with other hand-crafted features such as: word overlapping between the title and the article and the existence of highly polarized words from a lexicon (ex. fake, hoax). Their baseline achieved an FNC-1 score value of 75% and 45.4% value of Macro F1.

4 Approach and Results

The literature work on the FNC dataset showed that the best results are not obtained with a pure deep learning architecture, and simple BOW representations showed a good performance. In our approach, we combine n-grams, word embeddings and cue words to detect the stance of the title with respect to its article.

4.1 Preprocessing

Before building the feature representation, we perform a set of text preprocessing steps. In some articles we found links, hashtags, and user mentions (ex. @USER), so we remove them to make the text less biased. Similarly, we remove non-English and special characters.

4.2 Features

In our approach we combine simple feature representation to model the title-article tuples:

- **Cue words**: We employ a set of cue words categories that were used previously in (Bahuleyan and Vechtomova, 2017) to identify the stance of Twitter users towards rumor tweets. As Table 1 shows, the cue words categories are *Belief, Denial, Doubt, Report, Knowledge, Negation and Fake*. The Fake cue list is a combination of some words from FNC-1 baseline polarized words list and words from the original list. The provided set of cue words is quite small, therefore, we use Google News word2vec to expand it. For each word, we retrieve the most 5 similar words. As an example, for the word ”misinform”, we retrieved ”mislead ”,”misinforming”,”disinform”,”misinformation”, and ”demonize” as the most similar words.

Feature	Example Words
Belief	assume, believe, think, consider
Denial	refuse, reject, rebuff, oppose
Doubt	wonder, unsure, guess, doubt
Report	evidence, assert, told, claim
Knowledge	confirm, definitely, support
Negation	no, not, never, don’t, can’t
Fake	liar, false, rumor, hoax, debunk

Table 1: The cue words categories and examples.

- **Google News word2vec embedding**: For each title-article tuple, we measure the cosine similarity of the embedding of each sentence. Also, we use the full 300 length embedding vector for both the title and the article. The sentence embeddings is obtained by averaging its words embeddings. Previously in (Ghanem et al., 2018a), the authors showed that using the main sentence components (verbs, nouns, and adjectives) improved the detection accuracy of a plagiarism detection approach[3] rather than using the full sentence components. Therefore, we build these embeddings vectors using the main sentence components. Furthermore, we maintain the set of cue words that showed in the previous point.

- **FNC-1 features**: we use the same baseline feature set (see Section 3.3).

4.3 Experiments

In our experiments, we tested Support Vector Machines (SVM) (using each Linear and RBF kernels), Gradient Boost, Random Forest and Naive Bayes classifiers but the Neural Network (NN) showed better results[6]. Our NN architecture consists of two hidden layers with rectified linear unit (ReLU) activation function as non-linearity for the hidden layers, and Softmax activation function for the output layer. Also, we employed the

[3] For extracting the main sentence components, we used NLTK POS tagger: https://www.nltk.org/book/ch05.html.

[5] The stackLSTM is not one of the FNC-1 participated approaches, but it achieved state-of-the-art result.

[6] The Scikit-learn python package was used in our implementation

Systems	Macro-F1
Majority vote	0.210
FNC-1 baseline	0.454
Talos (Baird et al., 2017)	0.582
UCLMR (Riedel et al., 2017)	0.583
Athene (Hanselowski et al., 2017)	0.604
stackLSTM (Hanselowski et al., 2018)	0.609
Our approach	0.596
Cue words	0.250
Word2vec embeddings	0.488

Table 2: The Macro F1 score results of the participants in the FNC-1 challenge.[5]

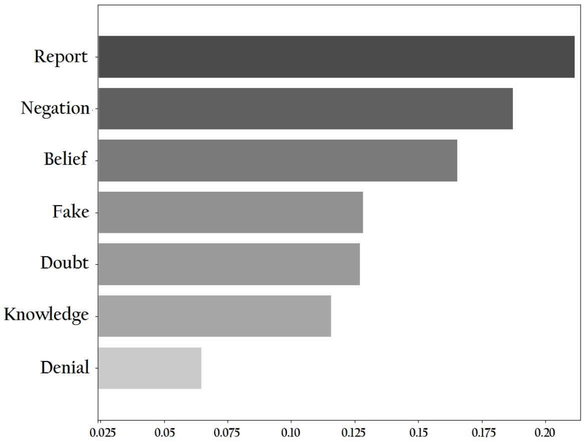

Figure 1: The importance of each cue words category using Information Gain.

Adam weight optimizer. The used batch size is 200. Table 2 shows the results of our approach and those of the FNC-1 participants. We investigated the score of each of our features independently. The word2vec embeddings feature set has achieved 0.488 Macro F1 value, while the cue words achieved 0.25. The extension of the cue words has improved the final result by 2.5%.

The tuples of the ”Unrelated” class had been created artificially by assigning articles from different documents. This abnormal distribution can affect the result of the cue words feature when we test it independently; since we extract the cue words feature from the articles part (without the titles) and some articles could be found with different class labels, this can bias the classification process. As we mentioned previously, the state-of-the-art result was obtained by an approach that combined LSTM with other features (see Section 2). Our approach achieved 0.596 value of Macro F1 score which is very close to the best result.

The combination of the cue words categories with the other features has improved the overall result. Each of them had impact in the classification process. In Figure 1, we show the importance of each category using the Information Gain. We extract it using Gradient Boost classifier as it achieves the highest result comparing to the other decision tree-based classifiers. The figure clarifies that *Report* is the category that has the highest importance in the classification process, where *Negation* and *Belief* categories have lower importance, whereas both of the *Denial* and *Knowledge* categories have the lowest importance. Surprisingly, both of the *Fake* and *Doubt* categories have a lower importance than the other three. Our intuition was that the *Fake* category will have the highest importance in discriminating the classes, where this category contains words that: may not appear in the ”Agree” class records, appear profusely in the ”Disagree” class (where the title is fake and the article proving that), and a medium appearance amount in the ”Discuss” class. Similarly, for the *Doubt* category, it seems that it may appear frequently in both ”Discuss” and ”Disagree” classes where its words normally mentioned when an article discusses a specific idea or when refuse it. To understand deeper our Information Gain results, we conducted another experiment to infer the importance of each category with respect to each classification class.

To do so, we use SVM classifier coefficients (linear kernel) to extract the most important category to each classification class. In our initial experiments, the SVM produced a result that is similar to the NN (58% Macro F1), so based on the good performance we used it in this experiment, where we couldn’t extract the feature importance using the NN. Once the SVM fits the data and creates a hyperplane that uses support vectors to maximize the distance between the classes, the importance of the features can be extracted based on the absolute size of the coefficients (vector coordinates). In Table 3 we show the importance of each category by their order. We can notice that for the ”Agree” class, generally, the categories that are used when there is a disagreement (Denial, Fake, Negation) tend to be less important than the other categories. On the contrary, for the ”Disagree”, disagreement categories appear in general in higher order comparing to the ”Agree” class.

#	Unrelated	Discuss	Disagree	Agree
1	Belief	Fake	Report	Belief
2	Negat.	Negat.	Fake	Report
3	Report	Belief	Denial	Doubt
4	Knowl.	Knowl.	Belief	Knowl.
5	Doubt	Denial	Negat.	Denial
6	Fake	Doubt	Knowl.	Fake
7	Denial	Report	Doubt	Negat.

Table 3: Importance order of the cue words categories for each class.

For the ”Discuss” class, due to the unclear stance towards the title where articles did not show a clear in favor or against stance, we can notice an overlapping in the highest order between the categories that are important for both ”Disagree” and ”Agree” classes. Finally, as we mentioned previously that the articles in the ”Unrelated” class are created artificially by assigning articles from different titles, the order of the categories is not meaningful.

5 Conclusion and Future Work

Fake news is still an open research topic. Further contributions are required, especially to deal automatically with the massive growth of information over the Web. Our work attempted to approach the stance detection of fake news using a simple model based on a combination of n-grams, word embeddings and lexical representation of cue words. These lexical cue words have been employed previously in the literature in rumors stance detection approaches. Although we used a simple feature set, we achieved similar results than the state of the art. This work is an initial step towards a further investigation of features to improve stance detection in fake news. As a future work, we plan to focus on summarizing the articles in the dataset. As we mentioned in Section 3.2, the length ratio difference between the titles and the articles is large. Therefore, summarizing the articles may be a worthy attempt to improve the comparison between the two text fragments.

Acknowledgement

This research work was done in the framework of the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

References

Hareesh Bahuleyan and Olga Vechtomova. 2017. Uwaterloo at semeval-2017 task 8: Detecting stance towards rumours with topic independent features. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 461–464.

Sean Baird, Doug Sibley, and Yuxi Pan. 2017. *Talos Targets Disinformation with Fake News Challenge Victory*. http://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.

Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims, task 2: Factuality. In *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Avignon, France. CEUR-WS.org.

Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 1353–1357. International World Wide Web Conferences Steering Committee.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*.

Omar Enayet and Samhaa R El-Beltagy. 2017. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 470–474.

Bilal Ghanem, Labib Arafeh, Paolo Rosso, and Fernando Sánchez-Vega. 2018a. Hyplag: Hybrid arabic text plagiarism detection system. In *International Conference on Applications of Natural Language to Information Systems*, pages 315–323. Springer.

Bilal Ghanem, Manuel Montes-y Gòmez, Francisco Rangel, and Paolo Rosso. 2018b. Upv-inaoe-autoritas - check that: An approach based on external sources to detect claims credibility. In *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Avignon, France. CEUR-WS.org.

Bilal Ghanem, Manuel Montes-y Gòmez, Francisco Rangel, and Paolo Rosso. 2018c. Upv-inaoe-autoritas - check that: Preliminary approach for checking worthiness of claims. In *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Avignon, France. CEUR-WS.org.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, and Felix Caspelherr. 2017. *Team Athene on the Fake News Challenge*. https://medium.com/@andre134679/team-athene-on-the-fake-news-/challenge-28a5cf5e017b.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance detection task. *arXiv preprint arXiv:1806.05180*.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1835–1838. ACM.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2018. We built a fake news & click-bait filter: What happened next will blow your mind! *arXiv preprint arXiv:1803.03786*.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. *arXiv preprint arXiv:1704.07221*.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *23rd International Conference on Applications of Natural Language to Information Systems*, pages 15–27. Springer.

Amr Magdy and Nayer Wanas. 2010. Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 103–110. ACM.

David M Markowitz and Jeffrey T Hancock. 2014. Linguistic traces of a scientific fraud: The case of diederik stapel. *PloS one*, 9(8):e105937.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 310–314.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Delip Rao and Dean Pomerleau. 2017. *Fake News Challenge*. http://www.fakenewschallenge.org.

Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *arXiv preprint arXiv:1707.03264*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Mariona Taulé, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, Viviana Patti, et al. 2017. Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*, volume 1881, pages 157–177. CEUR-WS.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 647–653.