A New Approach to Claim Check-Worthiness Prediction and Claim Verification

Shukrity Si
Jalpaiguri Govt. Engg College
India
sukriti.si98@gmail.com

Anisha Datta
Jalpaiguri Govt. Engg College
India
dattaanishadatta@gmail.com

Sudip Kumar Naskar
Jadavpur University
India
sudip.naskar@cse.jdvu.ac.in

Abstract

The more we are advancing towards a modern world, the more it opens the path to falsification in every aspect of life. Even in case of knowing the surrounding, common people can not judge the actual scenario as the promises, comments and opinions of the influential people at power keep changing every day. Therefore computationally determining the truthfulness of such claims and comments has a very important societal impact. This paper describes a unique method to extract check-worthy claims from the 2016 US presidential debates and verify the truthfulness of the check-worthy claims. We classify the claims for check-worthiness with our modified Tf-Idf model which is used in background training on fact-checking news articles (NBC News and Washington Post). We check the truthfulness of the claims by using POS, sentiment score and cosine similarity features.

1 Introduction

Today we live in a world where falsehood seems to reflect everywhere be it in administration, sports, entertainment sector and even in the education field. Many popular and influential personalities seem to be vulnerable in keeping their words. The opinions and comments they make, their claims keep changing frequently. Therefore we can not blindly rely on present news. During the 2016 US presidential campaign, people came to realize how fake news could be spread in mainstream news channels and social media. (Alexandre Bovet, 2019) reported the influence of fake news on social media during the election. They showed that about 171 million tweets were made during the election among which 25% were fake or extremely biased. Many journalists started investigation into identifying the actual truth. However, it was a time consuming and tedious task to do the work manually. This problem gave rise to the concept of automatic fact and claim checking. Research has been going on since then to effectively tackle this problem which has proved to be a very challenging problem. Typically for the claim verification task, relevant evidence related to the claims is collected first and then the claim is compared with the evidence to know the actual fact. (Giovanni Luca Ciampaglia and Flammini., 2015) did this with the help of knowledge graph taken from Wikipedia.

We propose a suitable rule based approach with the help of feature engineering for the task. Our work consists of two tasks, first we extract the claims which are check-worthy and then we verify the truthfulness of these check-worthy claims. We carried out our experiments on the dataset of the Fact Checking Master (Preslav Nakov and Martino., 2018) shared task, organized in CLEF-2018, which deals with fact checking on the U.S. Presidential debate articles of 2016.

Extraction of check-worthy claims is carried out in two processes i.e. supervised and unsupervised approach. In unsupervised approach, the check-worthy claims are extracted with POS tags and K-Means Clustering algorithm. Dataset related to claims can be generated from any conversation with the help of this method. In supervised approach, we have collected some fact checking news articles (NBC News and Washington post) for background training and a modified Tf-Idf model is created to classify the claims whether check-worthy or not. Cosine similarity, Sentiment scores and POS tags are also used here. On comparing with the original labels, this model gives an accuracy of 98.6% when passed along with GBM.

Claim verification is performed by comparing the classified check-worthy claims with the fact checking news articles to verify their truthfulness. POS tagging, Cosine similarity are used to search for the explanations of the claims from the arti-

Proceedings of the 17th International Conference on Natural Language Processing, pages 155–160
Patna, India, December 18 - 21, 2020.

cles. With all these features together we make a hypothesis for the final classification.

These two tasks are consecutively done in a work (Pepa Gencheva and Koychev, 2017) previously. Else there are many works to influence on individual approaches of the model. The paper is divided into many sections, section 2 describes related works, dataset in section 3, features and proposed methodology in section 4 and 5 respectively, results in section 6 and conclusion is given in section 7.

2 Related Works

Fact-Checking has become a trending topic recently. Zhou and Zafarani (2018) provides a survey on fake news research and their study focuses on fake news from four perspective – the false knowledge it carries, its writing style, its propagation patterns, and the credibility of its creators and spreaders. Pepa Gencheva and Koychev (2017) extracted the check-worthy claims by comparing them with some popular news articles and verified the truthfulness of the claims using Support Vector Machines (SVM) and Feed-Forward Neural Networks (FNN).

Mihai Surdeanu and Manning (2010) used Conditional Random Field (CRF) for legal claim identification. Firstly they used Optical Character Recognition (OCR) to convert PDF documents into text and then they used four types of CRF architectures for the actual task. Datta and Si (2020) reported a work on fake news identification in which sentiment scores and Tf-Idf are used as features to build a Majority Voting model with four classifiers - Gradient Boosting, Random Forest, Extra Tree and XGBoost classifiers to identify fake news. Ghanem et al. (2019) showed that false news has various emotional patterns to mislead the readers and with the emotional sentiment features they propose an LSTM model for the classification task.

Suzuki and Takatsuka. (2016) proposed a keyword-extraction model for verifying patent claims. RoyBar-Haim and Slonim performed claim stance classification by automatic expansion of the initial sentiment lexicon and by using SVM with unigrams. Naeemul Hassan and Tremayne (2015) developed a system called ClaimBuster which monitors Twitter and retweets the check-worthy factual claims it finds and produces true - false verdicts for these types of factual claims. Moin Nadeem (2019) reported an end to end fact-checking system, FAKTA, using document retrieval from various media sources, evidence extraction and linguistic analysis. Dieu-Thu Le and Blessing (2016) used Convolutional Neural Networks for the task. Rob Ennals and Rosario. (2010) developed another fact-checking system, DisputeFinder, which works on already verified claims. Ayush Patwari and Bagchi. (2017) used LDA topic modeling, POS tuples and Bag-of-Words as features and SVM is used for clustering. Wang. (2017) and Nicole OBrien and Boix. (2018), proposed different models to classify factuality of claims aimed at only input claims and their metadata.

3 Dataset

We carried out our experiments on the dataset of the Fact Checking Master (Preslav Nakov and Martino., 2018) shared task, organized in CLEF-2018. The dataset contains stated claims of the U.S. Presidential debates (2016) with a total of 1,403 sentences in the first Presidential debate and 1,303 sentences in the second Presidential debate. There are four speakers - Holt (host of the first Presidential debate), Cooper (host of the second Presidential debate), Hillary Clinton and Donald Trump. All the claims made by Mr. Trump and Ms. Clinton in these debates can be sensed manually since all of the claims show some actions happened in the past, or any comments or actions from the opponent in the past. Observing the patterns, we analysed the prominent features and developed our models.

4 Proposed Methodology

4.1 Claim Check-Worthiness Prediction

In the dataset, there are many statements which should be prioritized to be fact-checked as claims. We emply both supervised and unsupervised approach for this task.

4.1.1 Supervised Approach

We used modified Tf-Idf model with Gradient Boosting classifier for the work. Term Frequency (tf) measures how frequently a term occurs in a document or text. Since every document is different in length, it is possible that a term can appear more frequently in the longer documents than in the shorter ones. Thus, term frequency is normalized by the document length, i.e., the total number of terms in the document.

Idf is generally defined by the logarithm of the ratio of total number of documents in the dataset and the number of documents with that term in

them. The idf calculation is modified in our model. We have taken the statements from the data as documents in Tf calculation. The modified Idf is defined as the logarithm of the ratio of total no. of documents (the explanations from the fact-checking articles) and number of documents with the term in them. In case of normal Tf-Idf count, the documents (or, sentences) used in both Tf and Idf belong to the same article. But in our case, Tf count takes the actual data but Idf count takes the fact-checking news articles for consideration. The reason behind this modification is to check word similarity between a claim and an explanation. The more the similarity is, the more an explanation is related to the claim. This process helps in background knowledge training.

Modified $Idf =$

$$log_e \frac{Total\ no.\ of\ documents\ in\ the articles}{No.\ of\ documents\ with\ the\ term}$$

But in case, if the term is not present in the Fact-Checking articles, the Idf count is taken as 0. Now the Tf and Idf is multiplied to calculate the modified Tf-Idf count(feature model).

$$Modified\ Tf-Idf = Tf * Modified\ Idf \quad (1)$$

Therefore we modify Tf-Idf in this way to build a feature matrix which can help us to establish a relation between the background articles and the statements.

If the calculated Tf-Idf is a non-zero number, then the term is present in the background article. This increases the chance of the statement with the term to be classified as a check-worthy claim as it is related to one of the explanations present in the articles. As this is a supervised approach, we use the labels of our data. In the data, if the statement is a check-worthy claim then it is assigned to 1 or otherwise 0. The Tf-Idf feature model is now fitted into Gradient Boosting classifier for final classification.

4.1.2 Unsupervised Approach

In this approach, our main goal is to extract the check-worthy claims from the debate dataset without any labels. This approach can be used in future to create new dataset related to claims from any debates, conversations or interviews. Here we study the data very carefully to understand the features of check-worthy claims which are described below.

Features - Now an example of a claim is -
"Ford is leaving, you see that, their small car division leaving, Thousands of jobs leaving Michigan, leaving Ohio."
In this case, 'leave' verb is in continuous tense, and there are proper nouns like 'Ford', 'Michigan', 'Ohio'. So the statements containing proper nouns and continuous tense have a great chance to be check-worthy claims.
"He approved NAFTA, which is the single worst trade deal ever approved in this country."
This is a claim made by Trump. The adjective 'bad' is in superlative form in this sentence, verb 'approve' is in past tense and a connective word 'which' is used here. Now if any person uses other person's statements in indirect speech with a connective word, then there is a strong possibility that the sentence is a check-worthy claim. Because the person may change other's statement in his own way, and the statement should be checked whether it was actually stated or not. Now if anyone says - "Paolo Coelho said that he was the best writer of the world." This sentence is a claim indeed. Paolo Coelho might say that he is one of the best writer of the world, but the person distorted his statement. So this type of claims are check-worthy.

Therefore with all these features, we have made separate matrices and merged them all together. This merged feature matrix is passed along with a unsupervised machine learning algorithm called 'K-means Clustering' to create two clusters of check-worthy claims and non-claims.

We have used these two approaches for the classification. Among them the supervised approach is more suitable for our work, it gives better results than the other. But the unsupervised approach can help us to generate claim dataset without any labels. But this model needs further modification.

4.2 Claim Verification

After extracting the check-worthy claims from the dataset, we are now left with 17 and 16 claims for the 1st and 2nd presidential debate respectively. These claims now need to be checked for truthfulness by comparing with the existing fact-checking articles. The second part of the dataset contains labels of the claims according to their truthfulness. There are 3 labels as True, False and Half-True. So, the next task is to make a suitable model to classify the check-worthy claims. Now this work is divided into two parts, first part is the extraction

of related explanations for all the claims and the second part is to verify whether the claims are true or not by comparing with the explanations. We have used NBC news and Washington post fact-checking articles to get the true explanations of the check-worthy claims. They are described below.

4.2.1 Explanation Extraction

The first goal of this task is to find the proper explanations of the claims. We have used Cosine Similarity algorithm with the help of POS tags.

- POS tags - We compare each check-worthy claim with all the explanations given in the NBC news and Washington post articles. The first step is to POS tag each sentence. Some of the tags are given more importance than the others. These are - nouns (proper nouns), normal pronouns and possessive pronouns, verbs (past, continuous and participle tenses), adjectives, adverbs and connectives ('that', 'which', etc.). These tags increase the chance of getting similar sentences. Therefore, all these POS tags are taken into consideration and a single feature matrix is formed by merging them all together.

- Cosine Similarity - We compute cosine similarity between combined POS tagged list of each of the check-worthy claims and the combined POS tagged list of each explanation of the fact-checking articles. The maximum output gives out the true explanation. Now the true explanations are placed beside the claims to check for the truthfulness.

4.2.2 Truthfulness Detection

The next task is to compare each check-worthy claim with its explanation and verify whether the claim is true or not. For this, we use sentiment scores and then build a new hypothesis which is used in the classification task.

Sentiment Score - We use VADER model (Hutto, 2014) to get the sentiment scores. It calculates the positive, negative, neutral and compound sentiment polarity for any sentence (in English language). With the help of this model, we have calculated the Sentiment Scores of each sentence, separated out the compound scores to be more precise and then compared the scores of each claim with its explanation. The compound score is calculated as in Equation 2, where, $a = (positive + negative + neutral)$ and α is a constant, say, $t = sentiment(claim) - sentiment(explanation)$.

$$Compound\ score = \frac{a}{\sqrt{a^2 + \alpha}} \tag{2}$$

Now, this needs to be standardised further to get the threshold value for every label (True, False and Half-True). We have separated the claims from the original dataset according to their labels. Then the calculated t is placed accordingly, their means and medians are calculated with respect to the labels.

Observing the results, we have chosen threshold values for each label. The values are chosen as given below.

If $t \in [0.40, \infty)$, then the claim is True.

If $t \in [0.20, 0.40)$, then the claim is Half-True.

If $t \in [-\infty, 0.20)$, then the claim is False.

We determine the threshold by calculating mean and median of the variables, but if we look into the scores, we can understand that the threshold should work in right way. If any claim is false, generally its explanation will carry negative words like 'he didn't', 'it's a lie' etc. As we subtract the sentiments, for false claims, it becomes negative. And for true claims, the value is in positive range and for half true it lies in between them. This is the hypothesis we propose for claim verification.

This gives new labels for the claims. We have compared them with the existing labels and got reasonable scores for each label.

5 Results and Discussion

We calculate the values of accuracy, precision, recall and f-score for each model proposed in the paper. Confusion matrices are also shown for result visualization for both the debate articles. All the results are analysed below.

5.1 Claim Check-worthiness

We work on two dataset, 1st and 2nd US presidential candidate debate articles. We have used a modified version of Tf-Idf (a new method we have proposed here) model to extract check-worthy claims. We have tested the model with the original labels of the dataset with the help of GBM classifier. The results for both datasets are given in the table 1 and table 2.

Figure 1: Confusion Matrix of 1st debate article results

Figure 2: Confusion Matrix of 2nd debate article results

Table 1: Results on 1st-Presidential Debate-

Model	Accuracy	Precision	Recall	F-Score
GBM	98.65%	98.66%	98.64%	98.42%

Table 2: Results on 2nd-Presidential Debate-

Model	Accuracy	Precision	Recall	F-Score
GBM	99%	99.01%	99%	98.86%

Now we can see that the results of both debates are very good. The confusion matrices for the models on 1st and 2nd dataset are shown in figure 1 and 2 respectively.

Now there arise some cases of wrong predictions. A few check-worthy claims are present which cannot be extracted by our model but most of the predictions are correct. But there are no such extracted claims which are not originally check-worthy. Here we can conclude that our model is giving a very good performance for the check-worthiness problem. Although we will work on it for improvement.

5.2 Claim Verification

We have used some features like- POS tagging, Cosine Similarity and Sentiment Score on the extracted check-worthy claims to extract the true explanations from the fact-checking articles and verify the truthfulness of the claims. This approach on both the debate articles has brought good results for the labels - True and False. But for label Half-True, for some ambiguity, the result is not so good like the others. For this reason, the overall result has decreased. The individual results for each label and on a whole are shown below in the table 3, 4, 5, 6.

Results on 1st-Presidential Debate

Table 3:

Accuracy	Precision	Recall	F-score
64.705%	74.50%	64.70%	62.64%

Table 4:

	True	False	Half-True
Accuracy	75%	88%	33.3%

Results on 2nd-Presidential Debate

Table 5:

Accuracy	Precision	Recall	F-score
62.5%	67.05%	62.5%	61.25%

Table 6:

	True	False	Half-True
Accuracy	50%	77.7%	33.3%

All the predictions of our model is discussed and we realise that the models are facing some difficulties to classify the class half true. But if the half-true label is considered as false then the result is fine. But we are working on the improvement of this model for the particular class half true.

6 Conclusion and Future Work

The method which we use gives a very good result in check-worthiness classification, and the approach we propose here is a modified version of Tf-Idf for using background fact checking article. This can help us to do further research in background training. Other approach we have used is unsupervised approach, where we studied every feature very carefully and built a model. Though this model needs further modification but it can be used to generate new data on claims from any debate, conversation or interview. So it can contribute to the field of dataset creation. Now for Claim truthfulness, we first extracted the proper explanations from the articles, then applied our own threshold for classification. This approach works good for the class true and false, but for half true, it needs tuning and modification.

Further we can apply our modified Tf-Idf model in other researchers' works and check the efficiency. The model and hypothesis for truthfulness verification needs further improvement. We will try to add deep learning methods also. It can be concluded that our proposed model is very versatile and can be used in other fields as well.

References

Hernan A. Makse. Alexandre Bovet. 2019. Influence of fake news in twitter during the 2016 us presidential election. ArXiv:1803.08491v2 [cs.SI].

Dan Goldwasser Ayush Patwari and Saurabh Bagchi. 2017. Tathya:amulti-classifier system for detecting check-worthy statements in political debates. In *In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore, CIKM '17, pages 2259 to 2262.*

Anisha Datta and Shukrity Si. 2020. A supervised machine learning approach to fake news identification. In *Intelligent Data Communication Technologies and Internet of Things*, pages 197–204, Cham. Springer International Publishing.

Ngoc Thang Vu Dieu-Thu Le and Andre Blessing. 2016. Towards a text analysis system for political debates. on. LaTeCH.

Bilal Ghanem, Paolo Rosso, and Francisco M. Rangel Pardo. 2019. An emotional analysis of false information in social media and news articles. *ArXiv*, abs/1908.09951.

Luis M. Rocha-Johan Bollen Filippo Menczer Giovanni Luca Ciampaglia, Prashant Shiralkar and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. PLOS ONE 10(6):1 to 13.

E.E. Hutto, C.J. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. eighth international conference on weblogs and social media. *(ICWSM-14). Ann Arbor, MI, June 2014.*

Ramesh Nallapati Mihai Surdeanu and Christopher Manning. 2010. Legal claim identification: Information extraction with hierarchically labeled data. LREC.

Brian Xu Mitra Mohtarami-James Glass. Moin Nadeem, Wei Fang. 2019. Fakta: An automatic end-to-end fact checking system. In *Proceedings of NAACL-HLT 2019: Demonstrations, pages 78 to 83 Minneapolis, Minnesota, June 2 to June 7, 2019. Association for Computational Linguistics.*

Chengkai Li Naeemul Hassan and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *In Proceedings of the 24th ACM International Conference on Information and Knowledge Management. CIKM '15, pages 1835–1838.*

Georgios Evan-gelopoulos Nicole OBrien, Sophia Latessa and Xavier Boix. 2018. The language of fake news: Opening the black-box of deep learning based detectors. In *InProceedings of the Thirty-second Annual Conference on Neural Information Process-ing Systems (NeurIPS)–AI for Social Good.*

Lluıs M'arquez-Alberto Barron-Cede˜no Pepa Gencheva, Preslav Nakov and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. RANLP. Varna, Bulgaria, pages 267–276.

Tamer Elsayed Reem Suwaileh Lluıs M'arquez Wajdi Zaghouani Pepa Atanasova Spas Kyuchukov Preslav Nakov, Alberto Barron-Cede˜no and Giovanni Da San Martino. 2018. Overview of the clef 2018 checkthat!lab on automatic identfication and verification of political claims. In *In Proceedings of CLEF. Avignon, France, pages 372 to 387.*

John Mark Agosta Rob Ennals, Dan Byler and Barbara Rosario. 2010. What is disputed on the web?. In *In Proceedings of the 4th workshop on Information credibility. ACM, New York, NY, USA, WICOW '10, pages 67 to 74.*

Charles Jochim RoyBar-Haim, Lilach Edelstein and Noam Slonim. Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proceedings of the 4th Workshop on Argument Mining, pages 32–38 Copenhagen, Denmark, September 8, 2017. c 2017 Association for Computational Linguistics*.

Shoko Suzuki and Hiromichi Takatsuka. 2016. Extraction of keywords of novelties from patent claims. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1192–1200, Osaka, Japan, December 11 to 17 2016.*

William Yang Wang. 2017. iar, liar pants on fire":a new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume2: Short Papers), pages 422 to 426. Association for Computational Linguistics.*

Xinyi Zhou and Reza Zafarani. 2018. Fakenews: A survey of research, detection methods, and opportunities. ArXiv:1812.00315v1 [cs.CL].