Auto FC

Ryan Rice
ryanrice@stanford.edu
https://github.com/RyanRice02/Contextualized-Fact-Verification

Abstract

The internet – and social media in particular – has created a large volume of information that can be hard to verify. To address this problem, I seek to train an automated fact checker (Auto FC). The model makes use of Google's BERT, a pre-trained language model that has greatly advanced performance in several NLP tasks and is trained on a dataset constructed from the FEVER 1.0 dataset. While fact checking is not widely explored and the task has not been standardized, Auto FC achieves 89.2% test accuracy which outperforms every comparable model.

1 Introduction

Social media has given everyone the ability to create content for the world to consume. This is most important in the context of news, where it is referred to as citizen journalism. While there are merits to the democratization of news media, citizen journalism has brought about an unprecedented volume of information that at times is hard to trace back to its source and verify. Information quality would greatly improve if content could be checked automatically for validity, yet automated fact checking is not a popular and only minimally researched field of A.I.

In an attempt to solve to misinformation issue, Auto FC was trained to classify statements as either true, false, or indeterminable given a context passage that is considered ground-truth. In other words, the model takes as input a statement concatenated with a context, passes it through a neural network consisting of a pre-trained language model and an output layer, and outputs one of three prediction classes: supports, refutes, or not enough info.

The trained model achieves an accuracy of 89.2% which outperforms comparable models. Further analysis of the models results can be found in the Discussion section.

2 Related work

Automated fact checking is an under-researched natural language processing task. That being said, two similar tasks exist.

2.1 FEVER

The Fact Extraction and VERification contest was put on by a group of computer scientists from the University of Sheffield in July of 2018 [1]. The dataset provided had statements as inputs and evidence-classification pairs as outputs where the evidence was a span from the context and the classification was either true, false, or not enough info. The aforementioned context consisted of the English Wikipedia dump from 2017 and was used for all examples. The contest had two evaluation metrics: FEVER score and accuracy. Accuracy is solely calculated from the predicted class such that the percent accuracy is the percentage of examples properly classified. FEVER score is a combination of accuracy and evidence extraction where an example is only considered correct if it is properly

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

classified and the corrected supporting or refuting evidence is found – evidence is not reported for examples classified as not enough info.

In December of 2018, the authors of the dataset attempted to solve their own problem [5]. Their model consists of three components: a document retriever for gathering evidence, a sentence selector to sort gathered evidence by similarity to the input statement, and a recognizing textual entailment system for classifying statements based on the selected evidence. Given the large context, the document retriever, which pulls passages based on the cosine similarity of n-grams to the input statement, seems like a decent solution. From there, however, there seems to be a lot of hand-generated features where the model perhaps could have performed better using an end-to-end system. The model achieves 50.91% classification accuracy, ignoring evidence, and 86.16% classification accuracy given that the proper evidence was identified. This model is charged with a slightly different task than the model described in this paper; however, the 86.16% accuracy that is controlling for evidence is the most similar reported result found and was used as a baseline.

2.2 Fake News Challenge

The Fake News Challenge was put on during June of 2017 [2]. The task requires models to classify a headline and a corresponding article body as either in agreement, disagreement, discussion or unrelated. The evaluation metrics used are FNC score and relative score. The FNC score a weighted points metric where models are awarded 0.25 points for correctly classifying an input pairing as unrelated or related and another 0.75 points if given that the pairing is related, the correct agrees/disagree/discusses label is predicted. The relative score is the number of points earned as a percentage of the maximum score.

A pair of computer scientists, one from the University College London and the other from the University of Copenhagen, trained what they consider a hard to beat baseline for this task [4]. The model utilizes a multi-layer perceptron with one hidden layer and achieves 88.46% accuracy. The end-to-end approach taken seems very successful, and the authors report that it is on par if not better than the complex ensemble models with hand-crafted features it competes against.

3 Dataset and Features

The dataset used was constructed from the FEVER 1.0 dataset. Examples consist of a statement, a corresponding context passage, and a label denoting whether the statement is supported, refuted, or if there is not enough information to tell based on the context passage. The context passages are collected from a 2017 dump of the English Wikipedia.

The original FEVER task required models to use the entire Wikipedia dump as context and extract evidence supporting or refuting evidence for each statement. Because this is a somewhat different task, a significant amount of preprocessing was necessary to construct the dataset described above. Passages containing the evidence for statements that could be supported or refuted were matched to those statements. Statements where not enough information existed were paired with randomly selected passages, a technique borrowed from the baseline FEVER model.

Below is an example from the dataset.

Statement
Murda Beatz's real name is Marshall Mathers.

Context
Shane Lee Lindstrom (born February 11 , 1994), professionally known as Murda Beatz, is a Canadian hip hop record producer from Fort Erie, Ontario. He is noted for producing songs such as "No Shopping" by rapper French Montana, "Back on Road" by rapper Gucci Mane, Lindstrom has also produced several tracks for various artists such as Drake, Migos, Travis Scott and PartyNextDoor, among others.

Label
Refutes

The overall training set had 144,675 examples, the dev set had 9,943 examples, and the test set had 9,943 examples as well. Due to time and space constraints, a subset consisting of 10,000 training examples, 1,000 dev examples, and 1,000 test examples was used to train the model.

4 Methods

The model is trained by fine-tuning Google's BERT [3]. More specifically, the statement and context passage are concatenated into a single input, and this is fed into the base BERT model. From there, the final hidden state from BERT is passed through a single linear layer to generate predictions for each of the three classes: supported, refuted and not enough information.

BERT is a pre-trained language model architected from bidirectional Transformers. The base version of BERT – which is used for this model – consists of 12 Transformer blocks with hidden units sizes of 768 and 12 self-attention heads. The final hidden state of BERT is denoted as

$$H \in \mathbb{R}^{h \times 1}$$

where h is the number of hidden units. The final linear layer that is added to BERT is then applied to this final hidden state to produce output $L \in \mathbb{R}^{3 \times 1}$.

$$W \in \mathbb{R}^{3 \times h}$$

$$L = HW + b$$

A softmax is then applied to L to generate class predictions.

5 Experiments/Results/Discussion

5.1 Experiments

The model was trained for three epochs using a batch size of 32, maximum sequence lengths of 128, and a learning rate of 2e-5. These hyperparameters were selected because they were reported by Google as optimal for classification tasks using BERT. Due to the initial success of the model, a lack of resources, and extensive training times per epoch, a hyperparameter search was not conducted.

5.2 Results

The model was evaluated using accuracy as the primary metric, and 97% training accuracy and 89.2% test accuracy were achieved. These results outperform comparable models, and performance comparisons can be seen in the table below.

Model	Accuracy
Auto FC	**89.2**
FEVER	86.16
Fake News	88.46

5.3 Discussion

The disparity in training and testing accuracy suggests high variance, and I believe this could be corrected by using more of the available training data. Qualitative analysis of the model's performance shows some interesting results.

Confusion Matrix

True | Prediction	Supports	Refutes	Not Enough Info	Overall	Accuracy
Supports	286	44	0	330	86.67
Refutes	48	291	0	339	85.85
Not Enough Info	0	0	331	331	100
Overall	334	335	331	1000	89.2

As seen in the above matrix, the model has perfect precision and recall on examples labeled "Not Enough Info", and this performance suggests the model is extremely good at determining whether two pieces of text are related or not. This result isn't too surprising as the Fake News Challenge model reports a similar result for examples marked unrelated.

While not quite perfect performance, the model is still successful on the supports/refutes examples as well. Below is an example that the model correctly classified.

Statement
Fabian Nicieza has yet to work on a comic book.

Context
Fabian Nicieza (born 31 December 1961) is an Argentine-American comic book writer and editor who is best known for his work on Marvel titles such as X-Men, X-Force, New Warriors, Cable and Deadpool, and Thunderbolts, for all of which he helped create numerous characters.

Label
Refutes

Prediction
Refutes

The model's understanding that the statement is false – its language indirectly marks Fabian Nicieza as not a comic book writer while the context clearly describes him as one – is particularly promising. These results aren't always achieved, however, and there are still misclassified examples. Below is an example that the model got wrong.

Statement
Entourage (film) received generally positive reviews.

Context
Entourage is a 2015 American comedy film written, directed and co-produced by Doug Ellin. It serves as a continuation of the HBO TV series of the same name created by Ellin. It stars the principal cast of the show, Kevin Connolly, Adrian Grenier, Kevin Dillon, Jerry Ferrara and Jeremy Piven. The film was released on June 3, 2015, received generally negative reviews and grossed over $49 million.

Label
Refutes

Prediction
Supports

This example seems fairly straightforward, yet the model makes a mistake here. This type of error is not frequent and could likely be removed by reducing the model's variance with more training data.

6 Conclusion/Future Work

Overall, the model achieves promising results on a task that has been generally ignored. The model's performance compared to similar models speaks to the power of pre-trained language models for natural language tasks as well as the power of end-to-end systems.

Given more time and resources, the effects of more data and tuned hyperparameters would be explored. The biggest constraints in the training process were storage – the dataset was several gigabytes in size such that provided cloud computing storage exhausted in less than two weeks – and training time – the model took several hours per epoch on a reduced dataset such that training on the complete dataset would not have been feasible. With these constraints removed, the variance between the training and testing accuracies could almost certainly be reduced resulting in an even better model than the one trained.

References

[1] URL: http://fever.ai.

[2] URL: http://www.fakenewschallenge.org.

[3] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018.

[4] Benjamin Riedel et al. *A simple but tough-to-beat baseline for the Fake News Challenge stance detection task*. 2018.

[5] James Thorne et al. *FEVER: a large-scale dataset for Fact Extraction and VERification*. 2018.