WISH UPON IS A SUPERNATURAL HORROR THRILLER DOG.
+
WISH UPON IS A 2017 SUPERNATURAL HORROR THRILLER FILM DIRECTED BY JOHN R. LEONETTI...

AUTO FC

RYAN RICE | RYANRICE@STANFORD.EDU

WISH UPON IS A SUPERNATURAL HORROR THRILLER DOG.
=
FALSE

THE PROBLEM

WITH MORE CONTENT THAN EVER ON THE INTERNET, IT CAN BE HARD TO TELL WHAT IS TRUE AND WHAT IS NOT.

IN ORDER TO AID WITH FACT VERIFICATION, I LOOKED TO BUILD AN AUTOMATED FACT CHECKER (AUTO FC). THE MODEL FINE-TUNES GOOGLE'S BERT FOR SEQUENCE CLASSIFICATION IN ORDER TO PREDICT STATEMENTS AS TRUE, FALSE, OR NOT ENOUGH INFO GIVEN A CONTEXT PASSAGE.

THE MODEL ACHIEVED 89.2% TEST ACCURACY, OUTPERFORMING COMPARABLE SYSTEMS.

THE DATA

THE DATA USED WAS CONSTRUCTED FROM THE FEVER DATASET [1].

STATEMENTS WERE MATCHED WITH CONTEXT PASSAGES FROM THE ENGLISH WIKIPEDIA AND GIVEN TRUE/FALSE/NOT ENOUGH INFO LABELS.

THE DATA WAS SPLIT INTO A TRAINING SET OF 10000 EXAMPLES AND DEV/TEST SETS OF 1000 EXAMPLES EACH.

THE MODEL

THE MODEL CONSISTED OF FINE-TUNING THE BASE BERT MODEL FOR SEQUENCE CLASSIFICATION [3].

INPUTS WERE CONSTRUCTED BY CONCATENATING THE STATEMENT AND THE CONTEXT AND TRUNCATING TO A MAXIMUM SEQUENCE LENGTH OF 128.

A SINGLE LINEAR LAYER WAS ADDED TO THE BERT MODEL IN ORDER TO PREDICT ONE OF THREE LABELS — TRUE, FALSE, NOT ENOUGH INFO — FOR THE STATEMENT-CONTEXT INPUT PAIR.

THE RESULTS

GIVEN THE LIMITED RESOURCES, LIMITED TIME, AND IMMEDIATE SUCCESS OF THE MODEL, IT WAS ONLY TRAINED ONCE WITH A BATCH SIZE OF 32, A LEARNING RATE OF 2E-5, AND A MAXIMUM SEQUENCE LENGTH OF 128.

THE MODEL ACHIEVED 97% TRAINING ACCURACY AND 89.2% TEST ACCURACY.

DISCUSSION

TWO INTERESTING RESULTS FROM THE MODEL ARE ITS HANDLING OF INPUTS LABELED "NOT ENOUGH INFO" AND ITS COMPARATIVE ACCURACY.

THE MODEL MET THE FORMER WITH EXTREME SUCCESS, ACHEIVING 100% PRECISION AND RECALL ON THESE EXAMPLES. FOR THE LATTER, ITS ACCURACY WAS COMPARED TO THE BEST RESULTS FROM TWO FACT CHECKING TASKS, FEVER (86.16%) AND THE FAKE NEWS CHALLENGE (88.46%), AND IT OUTPERFORMS BOTH SYSTEMS [1][2]. IN FAIRNESS TO THE OTHER MODELS, THESE COMPARISONS AREN'T PERFECT AS THE TASKS AREN'T EXACTLY THE SAME. THIS LACK OF STANDARDIZATION IS WHY A STATE-OF-THE-ART MODEL HAS NOT BEEN ESTABLISHED.

FUTURE

GIVEN MORE TIME — AND POTENTIALLY RESOURCES — I WOULD TRAIN ON MORE DATA. NOT ONLY WAS TRAINING EXTREMELY TIME INTENSIVE, BUT THE DATA AND MODEL WERE EXTREMELY LARGE WHICH CAUSED ME TO QUICKLY RUN OUT OF STORAGE, THUS LIMITING THE AMOUNT OF TRAINING THAT COULD BE DONE. TO COPE WITH THIS, ONLY 10% OF THE POTENTIALLY AVAILABLE DATA WAS USED, BUT I THINK USING MORE OF THIS DATA WOULD IMPROVE THE MODEL BY DECREASING THE VARIANCE BETWEEN THE TRAINING AND TESTING ACCURACIES.

[1] JAMES THORNE ET AL.
FEVER: A LARGE-SCALE DATASET FOR FACT EXTRACTION AND VERIFICATION. 2018.

[2] BENJAMIN RIEDEL ET AL.
A SIMPLE BUT TOUGH-TO-BEAT BASELINE FOR THE FAKE NEWS CHALLENGE STANCE DETECTION TASK. 2018.

[3] JACOB DEVLIN ET AL.
BERT: PRE-TRAINING OF DEEP BIDIRECTIONAL TRANSFORMERS FOR LANGUAGE UNDERSTANDING. 2018.