Fact Extraction and Automated Claim Verification

Elizabeth Soper, Rohit Lalchand Vishwakarma, Swapnil Kishore

INTRODUCTION

Fact Extraction and Verification (FEVER)

Given a claim, find relevant evidence to support or refute it.

Critical problem for verifying information in an era of ‘fake news’

DATASET DESCRIPTION

FEVER dataset:

185,445 manually labelled claims

Labels: **‘SUPPORTED’**, **‘REFUTED’**, or **‘NOTENOUGHINFO’**

145,449 training examples

19,998 development

20,000 testing

DATASET DESCRIPTION

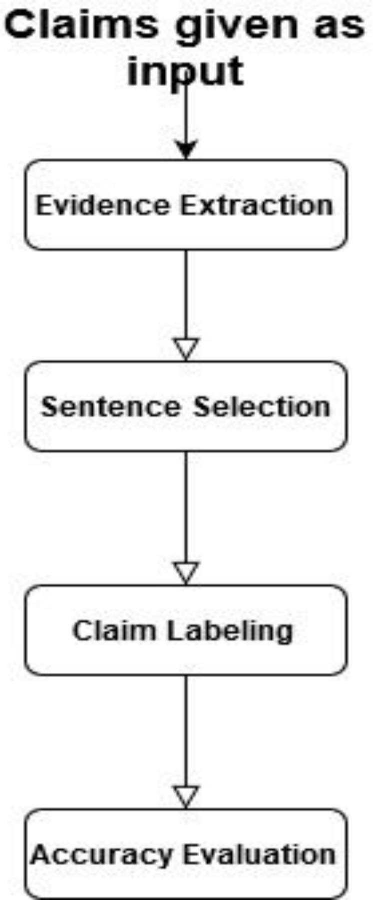
Format:

1. id: the id of the claim
2. label: One of {SUPPORTED, REFUTED, NOTENOUGHINFO}
3. claim: The text of the claim
4. evidence: A list of sentences relevant to the claim. (document id, index of sentence in document)

DATASET DESCRIPTION

To limit the scope of the task, evidence must come from a predefined corpus: over 5 million pre-processed Wikipedia pages from a June 2017 dump.

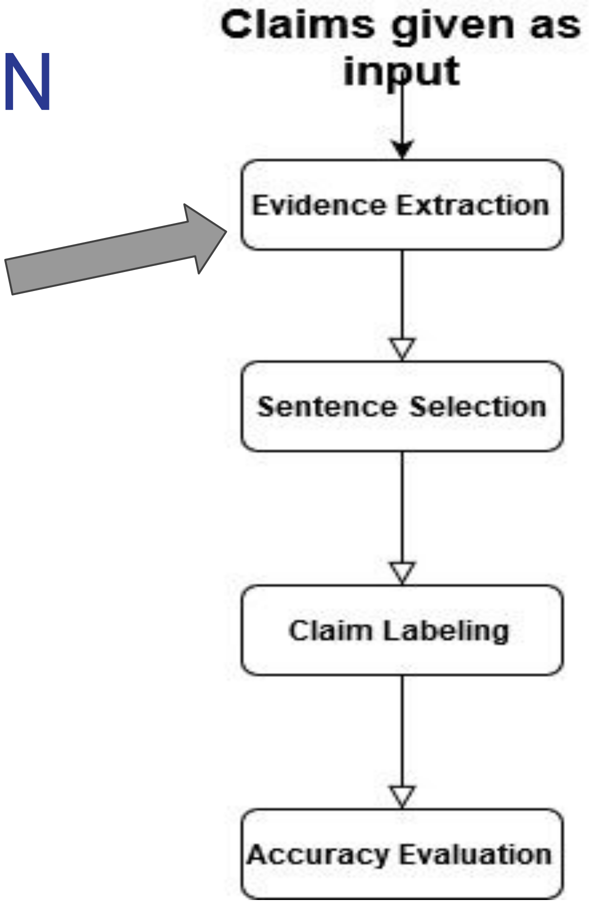
TASK DESCRIPTION



TASK DESCRIPTION

Step 1:

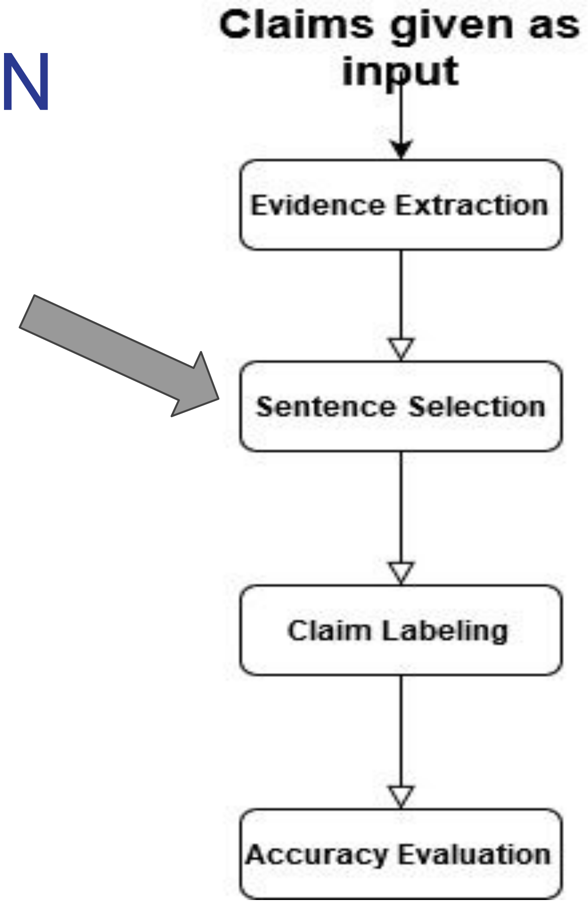
Within the corpus of Wikipedia documents, retrieve the ones most relevant to the claim



TASK DESCRIPTION

Step 2:

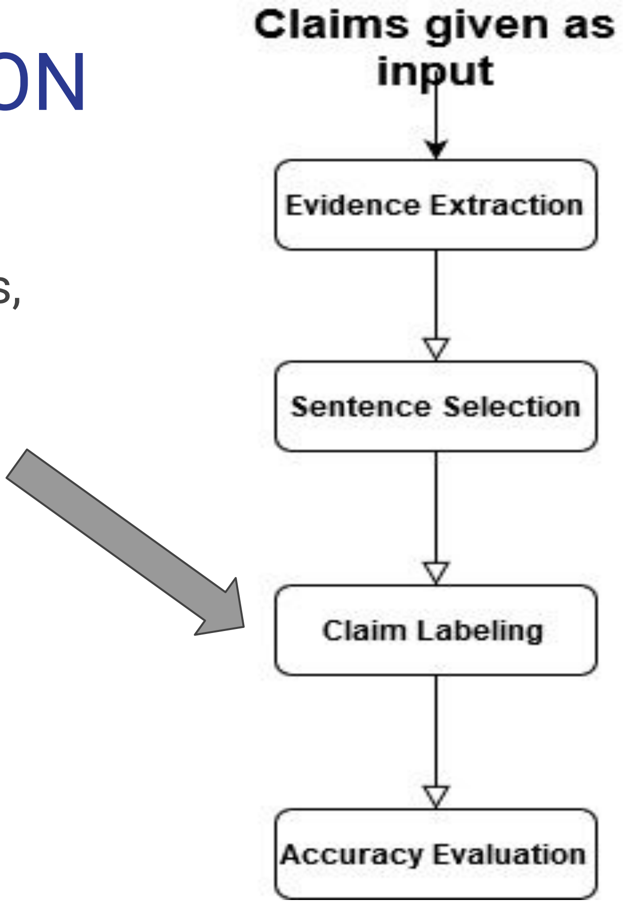
From the selected documents, choose the sentences most relevant to the claim.



TASK DESCRIPTION

Step 3:

Given the selected sentences, classify the claim as ‘Supported’, ‘Refuted,’ or ‘Not Enough Info’



Step 1: EVIDENCE EXTRACTION

1) **MediaWiki API**: ‘wikipedia’ library available in Python, used for Wikipedia scraping.
    a) Full-text: pass the full-text of the claim as input
    b) Phrases: extract the noun-phrases from the claim (using Python’s nltk) and pass only these noun-phrases as input.
    c) Keywords: use **RAKE** (Rapid Automatic **Keyword Extraction)** algorithm, which selects key phrases in text by analyzing the frequency of word appearance and its co-occurrence with other words in the text.

Step 1: EVIDENCE EXTRACTION

2) **Sequence Matcher:** use Python’s inbuilt sequence matcher to find the most relevant Wikipedia pages.

a) Compare title of each Wikipedia page to claim text and generate a score for each Wikipedia page.
b) Select three highest scoring pages.
c) Given the size of the corpus (>5 million documents), this method is extremely time-consuming and requires extensive resources. For this reason thorough *evaluation was not possible with this method.*

Step 2: SENTENCE SELECTION

1) **Raw TF-IDF Score:**

   rank the sentences of selected documents according to TF-IDF score

   the top scoring sentence from each source is selected

Step 2: SENTENCE SELECTION

2) TF-IDF Similarity Score:

Calculate the raw TF-IDF score for each sentence (as before)

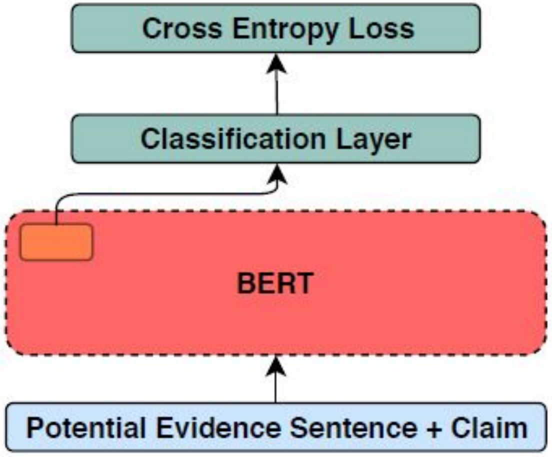
Calculate the cosine-similarity between the TF-IDF vectors of the claim and candidate evidence (sklearn.feature_extraction.text.TfidfVectorizer)

Rank candidate sentences according to the sum of both scores

Step 3: CLAIM LABELLING

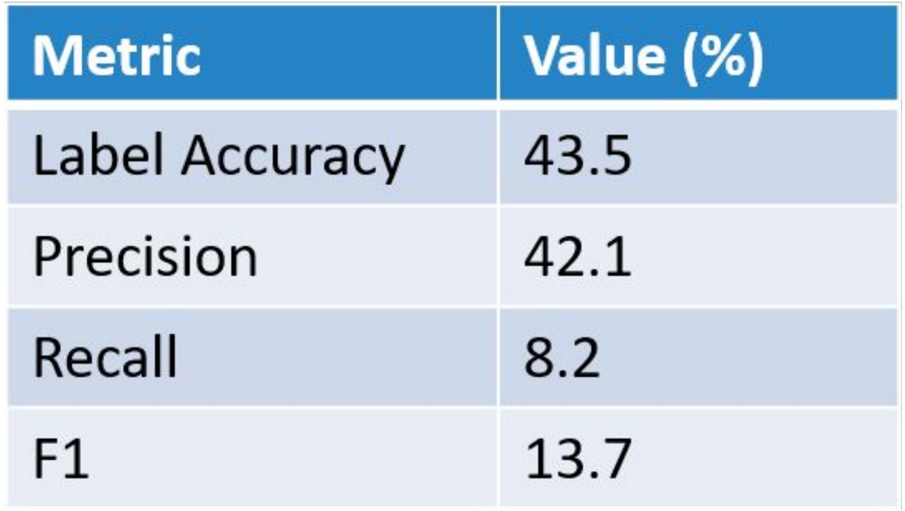
Use pre-trained BERT model to create representations for each claim and selected evidence sentences.

Feed BERT representations to a logistic regression classifier to generate a label (‘Supports,’ ‘Refutes,’ ‘Not Enough Info’)



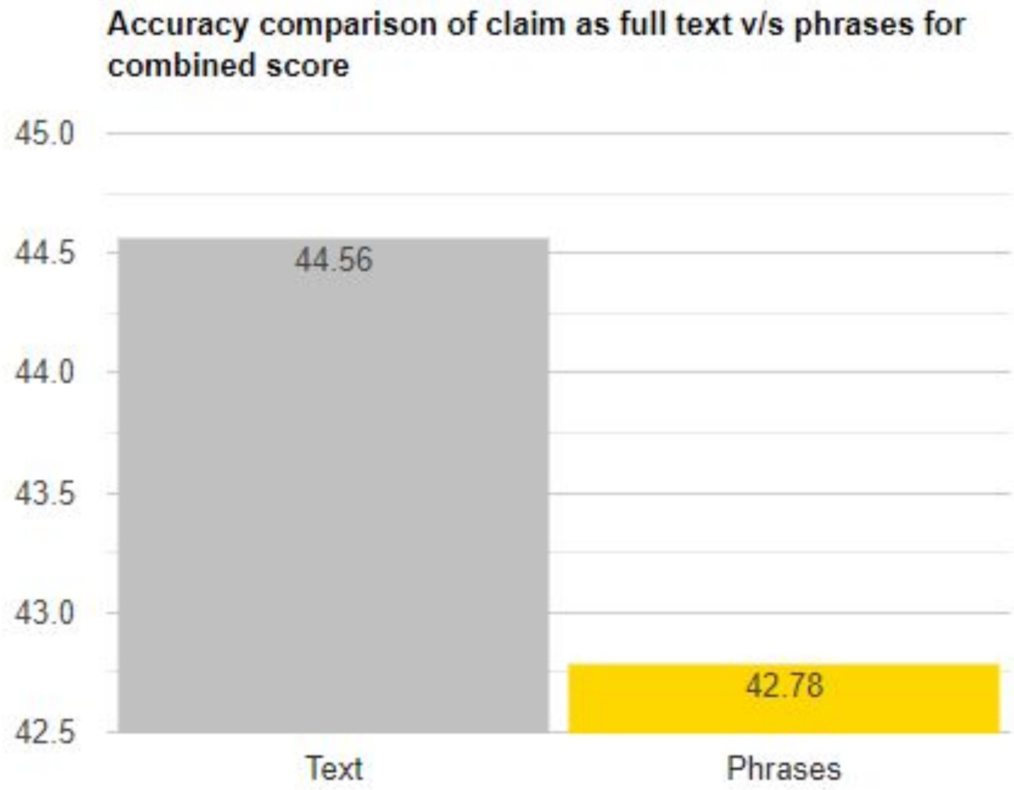
RESULT ANALYSIS: Baseline system

(using raw TF-IDF score to rank sentences)

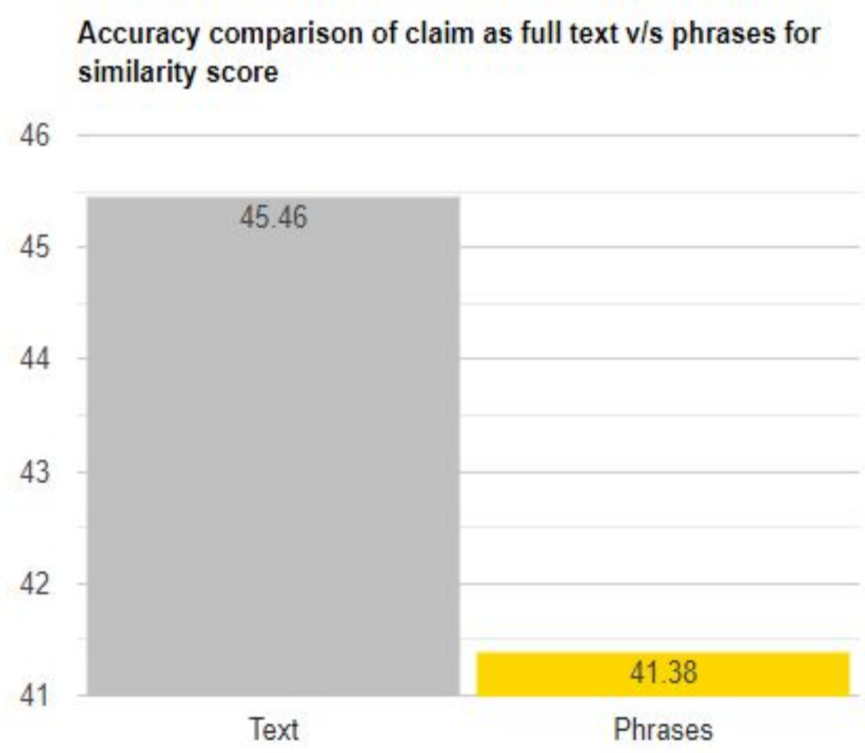


Metric	Value (%)
Label Accuracy	43.5
Precision	42.1
Recall	8.2
F1	13.7

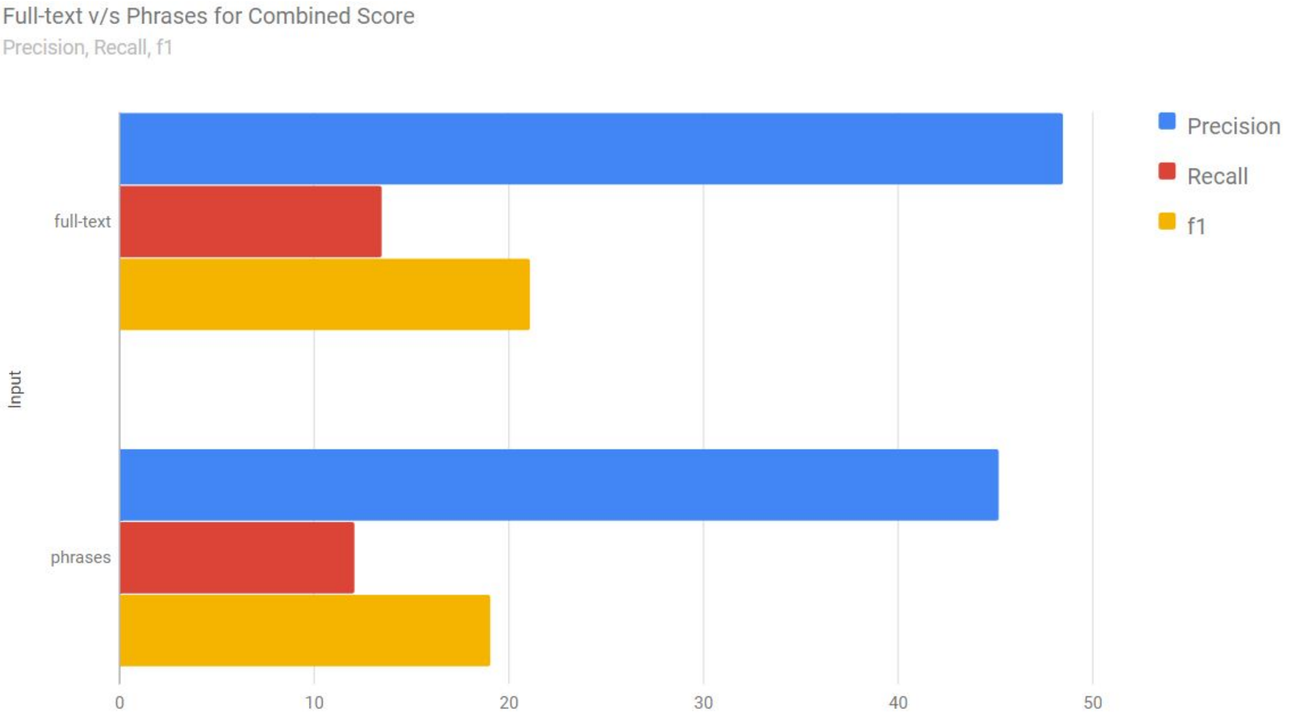
RESULT ANALYSIS



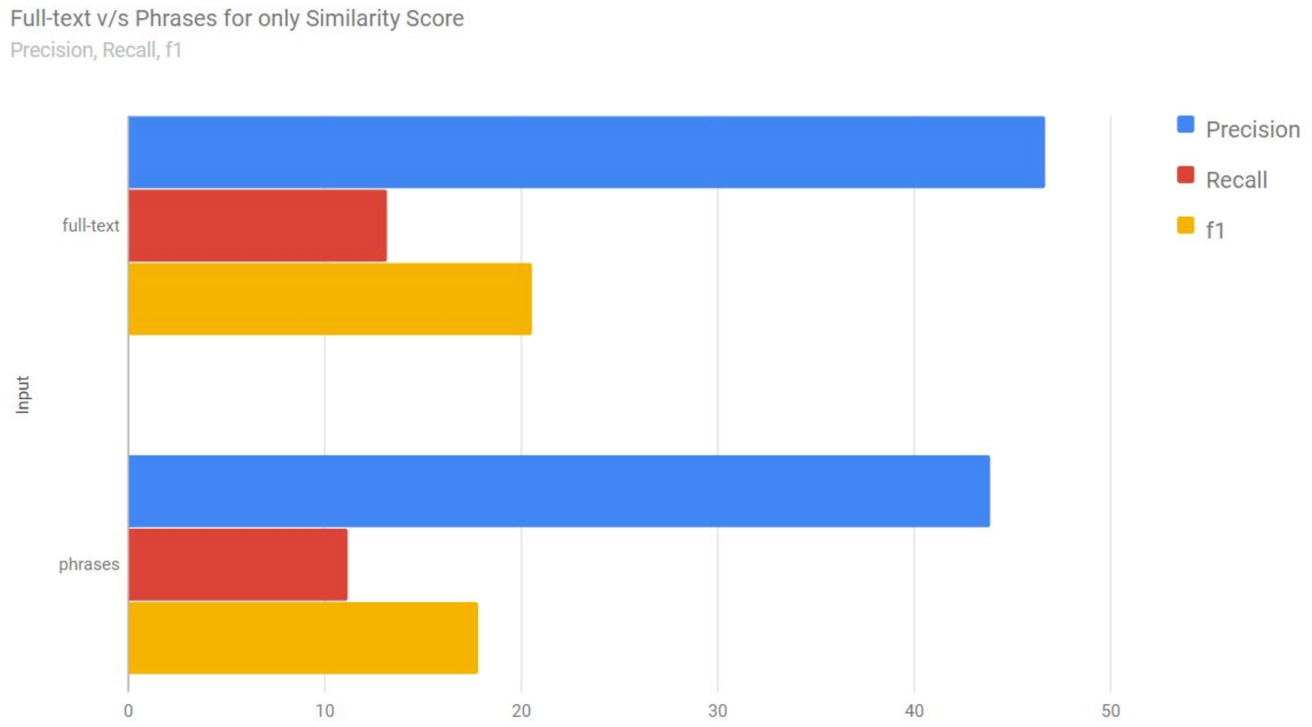
RESULT ANALYSIS


Accuracy comparison of claim as full text v/s phrases for similarity score

RESULT ANALYSIS

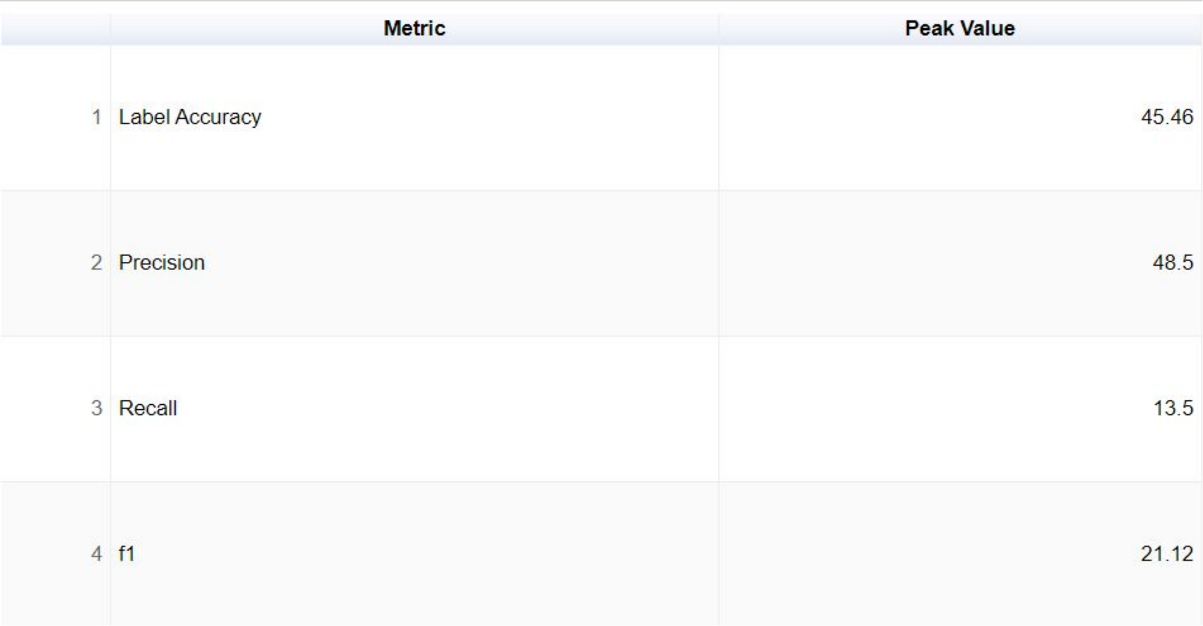


RESULT ANALYSIS



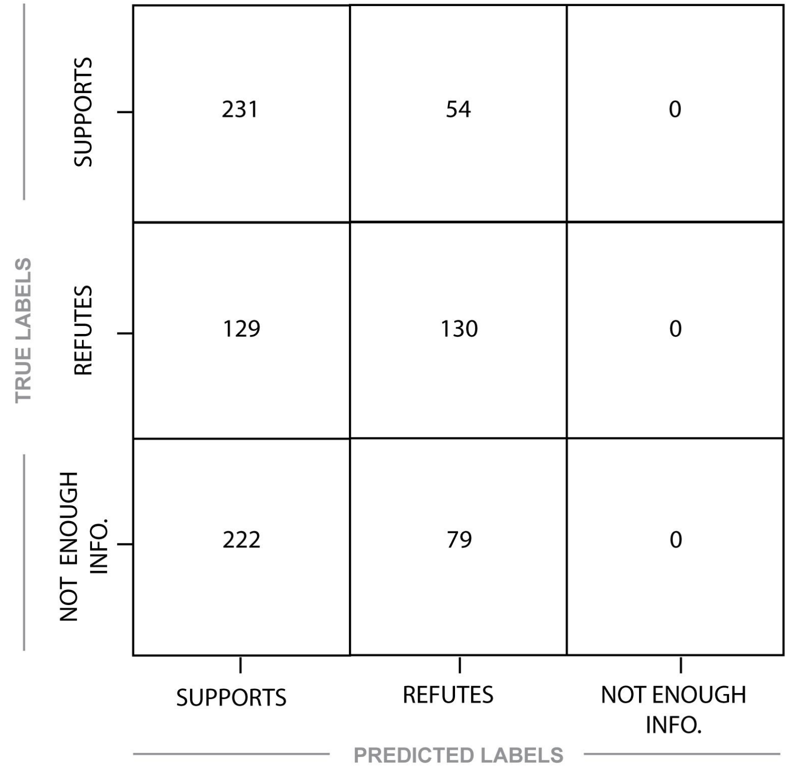
RESULT ANALYSIS: Final system

Here we have detailed the peak performance achieved by our system for each metric. (using cosine similarity for sentence selection)



	Metric	Peak Value
1	Label Accuracy	45.46
2	Precision	48.5
3	Recall	13.5
4	f1	21.12

RESULT ANALYSIS: Final system



TRUE LABELS \ PREDICTED LABELS	SUPPORTS	REFUTES	NOT ENOUGH INFO.
SUPPORTS	231	54	0
REFUTES	129	130	0
NOT ENOUGH INFO.	222	79	0

81.1% accuracy on SUPPORTS claims

50.2% accuracy on REFUTES

0% accuracy on NOT ENOUGH INFO

RESULT ANALYSIS

Performance breakdown by component:

Evidence extraction: retrieved at least one of relevant documents for 17.8% of verifiable claims

Claim Classification: 76.6% accuracy, given correct sentences

ERROR ANALYSIS: precision vs recall

Our system is stronger on precision than recall.

This is good for our task, as precision is the most important metric in fake news identification.

If the system says ‘true,’ we should be very confident it is actually true.

Goal is to prevent misinformation.

ERROR ANALYSIS

Example of correctly labelled claim:

{"claim": "A Milli is a song created by a recording artist who works in the genre of hip hop.", "sources": [["1990s in music", "15"], ["1990s in music", "11"], ["Hip hop", "19"]], "evidences": ["15\tA 2010 European survey conducted by the digital broadcaster Music Choice , interviewing over 11,000 participants , rated the 1990s as the second best tune decade in the last 50 years , while participants of an American land line survey rated the 1990s quite low , with only 8 % declaring it as best decade in music .\tMusic Choice\tMusic Choice\n", "11\tThe decade also featured the rise of contemporary country music as a major genre , which had started in the 1980s .\tcontemporary country music\tcountry music\n", "19\tRonald Savage , known by the nickname Bee-Stinger , who was a former member of the Zulu Nation , carved the term `` Six elements of the Hip Hop Movement " .\tZulu Nation\tUniversal Zulu Nation\tRonald Savage\tRonald Savage\n"]}

PREDICTED LABEL: SUPPORTS CORRECT LABEL: SUPPORTS

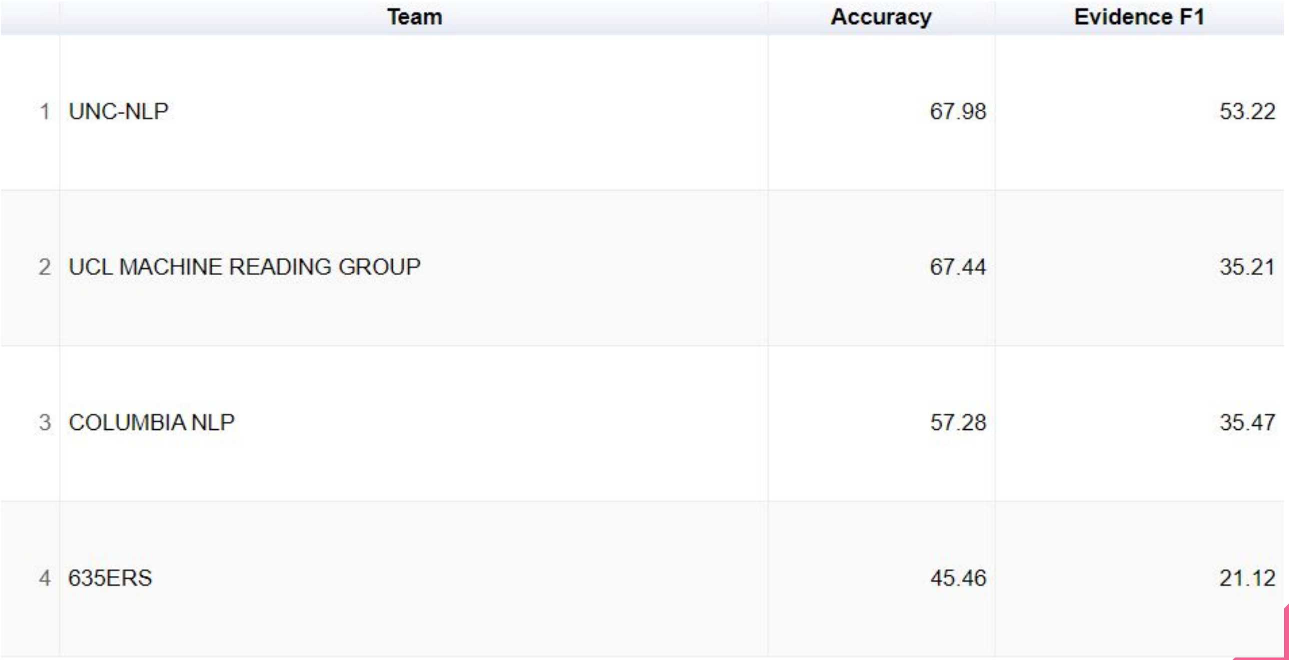
Example of incorrectly labelled claim:

{"claim": "Savages was exclusively a German film.", "sources": [["Nazi Germany", 19], ["List of Walt Disney Pictures films", 8], ["Africa Addio", 2]], "evidence": ["23\tChristian churches were also oppressed , with many leaders imprisoned .", "; unless they are credited as co-production partners -RRB- nor any direct-to-video releases , TV films , theatrical re-releases , or films originally released by other non-Disney studios .", "Africa\tAfrica\n2\tThe film was shot over a period of three years by Gualtiero Jacopetti and Franco Prosperi , two Italian filmmakers who had gained fame -LRB- along with co-director Paolo Cavara -RRB- as the directors of Mondo Cane in 1962 ."]}

PREDICTED LABEL: REFUTES CORRECT LABEL: SUPPORTS

NOTABLE WORK AND RESULTS

Comparison of our results to the top submissions from the 2018 Fever shared task:



	Team	Accuracy	Evidence F1
1	UNC-NLP	67.98	53.22
2	UCL MACHINE READING GROUP	67.44	35.21
3	COLUMBIA NLP	57.28	35.47
4	635ERS	45.46	21.12

DIRECTIONS FOR FUTURE WORK

1) Document retrieval:
   a) Search the FEVER corpus directly instead of using MediaWiki API, ranking documents according to similarity to claim.
   b) Named Entity Recognition instead of extracting noun phrases from the claim.
2) Sentence selection:
   a) Define a threshold for relevance. If no sentence is above threshold, claim is automatically labelled ‘Not enough info’
3) Claim classification:
   a) Train classifier on larger dataset.