Towards Automated Fact Checking

- Prabhat Agarwal (13CS10060)
- Priyank Palod (13CS30046)

Under the guidance of
Dr. Pawan Goyal and
Dr. Saurabh Bagchi

Fact Checking: Why?

Donald Trump

Republican presidential candidate

The unemployment rate may be as high as "42 percent."

in a press conference – Monday, September 28, 2015

POLITIFACT

Hillary Clinton

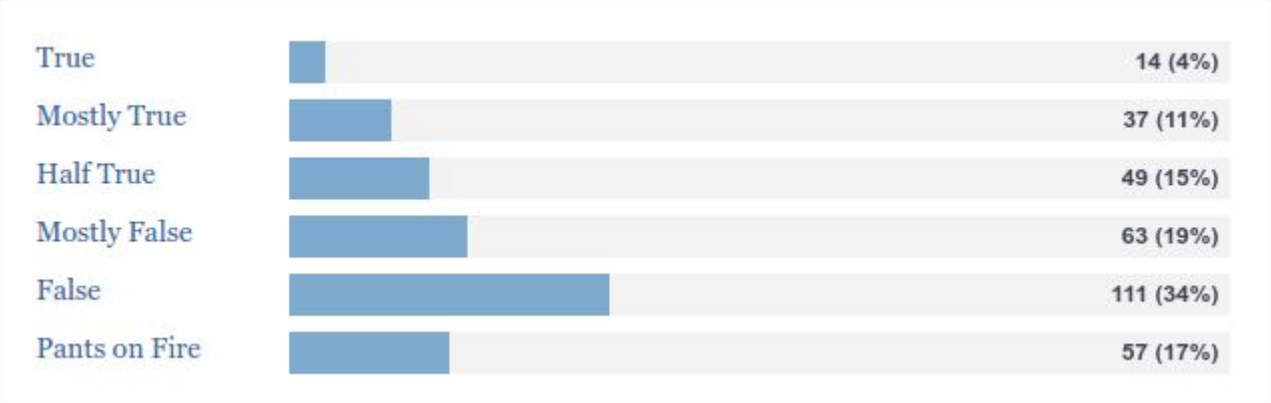
Democratic presidential candidate

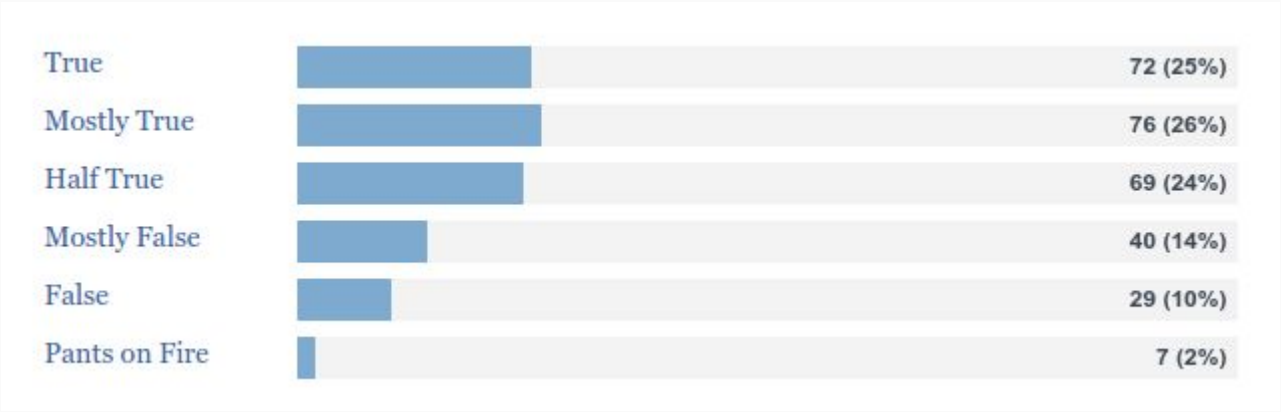
"It was allowed," referring to her email practices.

in an ABC interview – Thursday, May 26, 2016

POLITIFACT

Fact Checking: Why?





Source: www.politifact.com

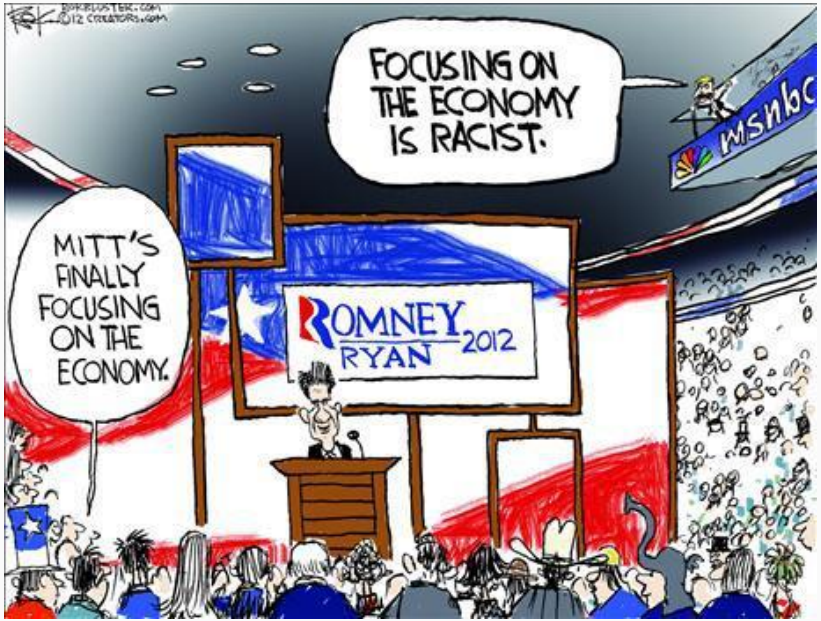
Why Automation?

Fact-checking presidential debate 'exhausting' says Toronto Star's Daniel Dale - CBCRadio

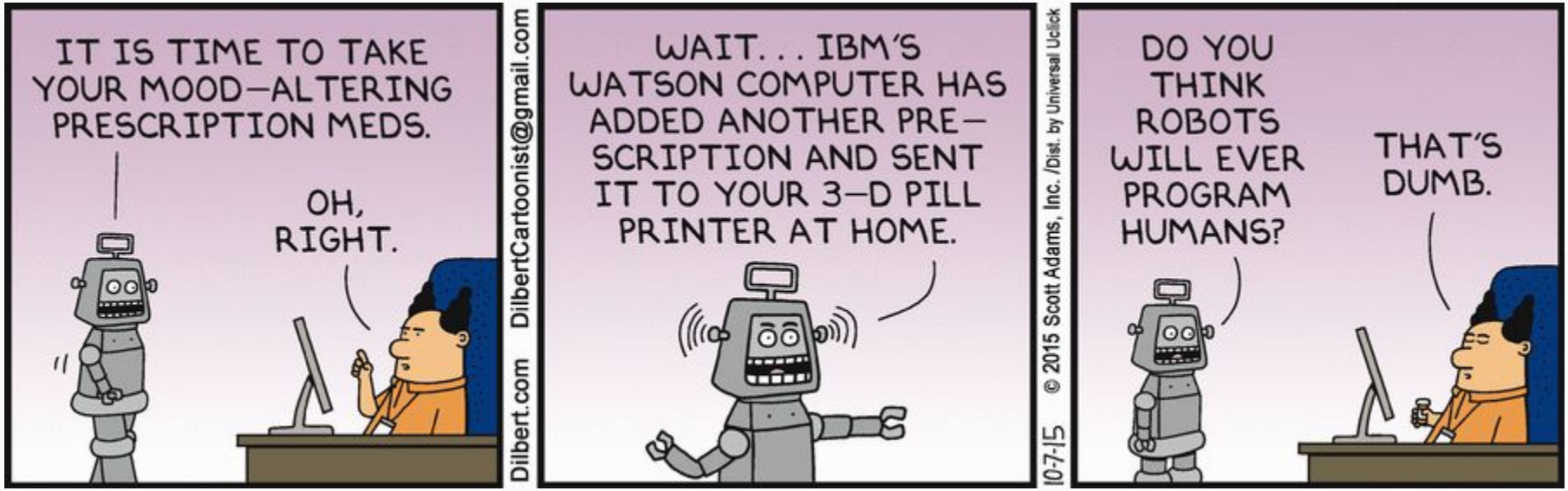
Why Automation?

Eight examples where ‘fact-checking’ became opinion journalism

- Washington Times



Why Automation?



Challenges

The Challenges : Why so hard?

- World Knowledge and Context
- Important details missing.
- Complicated Analysis
- Inherent Ambiguity in language
- Deliberate deception

"The Real household disposable income is rising."

"Under Donald Trump's tax plan, 51 percent of single parents would see their taxes go up."

"My grandfather immigrated to America."

"Thousands of Americans have been killed by illegal immigrants."

Broad Approach

Broad Approach

Claim Extraction → Source Identification → Information Analysis → Verdict with Arguments

Given a raw piece of text, identify and extract factual and check-worthy claims. Build a structural representation of the claim.

Identify sources of information about the claim, corroborates data from multiple sources with regards to quality and completeness.

Check the stance of the claims with the data from identified sources and previously checked facts.

Decide the truthfulness of the claim along with arguments and references.

Broad Approach

Claim Extraction	Source Identification	Information Analysis	Verdict with Arguments
Given a raw piece of text, identify and extract factual and check-worthy claims. Build a structural representation of the claim.	Identify sources of information about the claim, corroborates data from multiple sources with regards to quality and completeness.	Check the stance of the claims with the data from identified sources and previously checked facts.	Decide the truthfulness of the claim along with arguments and references.

Broad Approach

Claim Extraction

Given a raw piece of text, identify and extract factual and check-worthy claims. Build a structural representation of the claim.

Source Identification

Identify sources of information about the claim, corroborates data from multiple sources with regards to quality and completeness.

Information Analysis

Check the stance of the claims with the data from identified sources and previously checked facts.

Verdict with Arguments

Decide the truthfulness of the claim along with arguments and references.

Broad Approach

Claim Extraction

Given a raw piece of text, identify and extract factual and check-worthy claims. Build a structural representation of the claim.

Source Identification

Identify sources of information about the claim, corroborates data from multiple sources with regards to quality and completeness.

Information Analysis

Check the stance of the claims with the data from identified sources and previously checked facts.

Verdict with Arguments

Decide the truthfulness of the claim along with arguments and references.

Claim Extraction

The first step

We have worked on Claim Extraction. Given the transcript of a debate, our system returns all the check-worthy claims in a ranked order.

ClaimBuster automated live fact-checking

Find factual claims in your own text

2016 U.S. Presidential Debates

Hansard: Parliament of Australia

Tweets by @ClaimBusterTM

ClaimBuster Retweeted

West Wing Reports @WestWingReport

Electoral college update at 10:40 ET
Trump 167 (103 more needed)
Clinton 122 (148 more needed) (Virginia now hers)

11m

ClaimBuster Retweeted

West Wing Reports @WestWingReport

Electoral college update at 10:25 ET
Trump 167 (103 more needed) (Ohio now his)
Clinton 109 (161 more needed)

Dataset

Dataset Construction

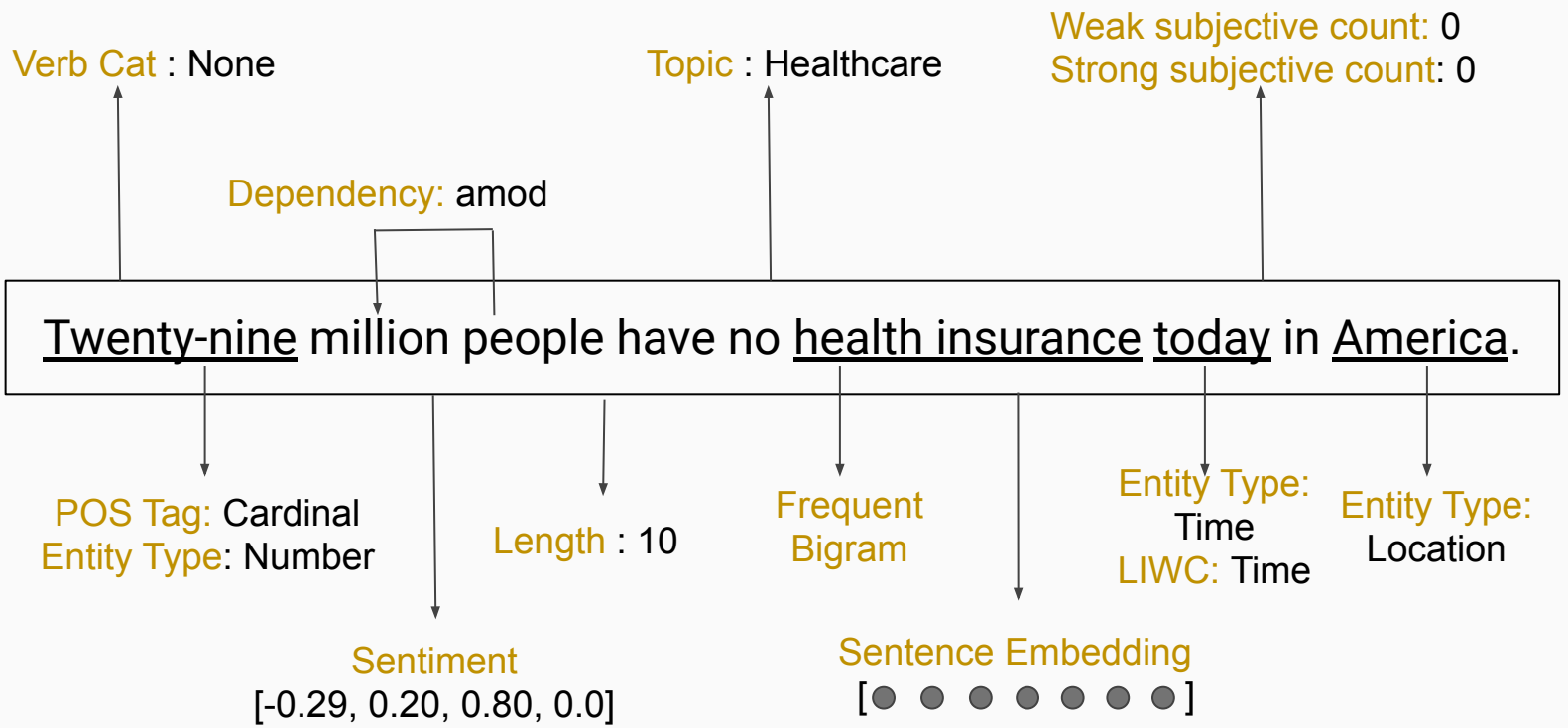
- Transcripts of 16 American Presidential debates in the primary elections 2016 (Aug 15 to April 16).
- 15235 statements by candidates: removed sentences not of any candidate.
- A sentence is considered check-worthy if it has been fact-checked by at least one of the organizations (Politifact, Factcheck, New York Times, Washington Post, NBC News, etc.)
- Manually annotated each sentence by referring to all the above organization.
- Finally had 608 sentences labelled check-worthy and 14627 sentences labelled not check-worthy.

Dataset - Errors

- We did an error analysis of the negative class by taking a small random sample and judging if the statement is check-worthy.
- Out of the 50 statements we took, 5 seemed check-worthy. Hence, negative class seems to have a 10% noise.
- Some statements are not checked because they have already been fact checked in some previous debate (2 statements or around 4%).
- Often, facts are not checked because of time/labor required. Fact-check websites are in a rush to send out the accurate checks in quick time.

Features

Feature Extraction



Results and Comparison

Results: Classification

- **Train Set**: 70% split, 426 labeled check-worthy
- **Test Set**: 30% split, 182 labeled check-worthy

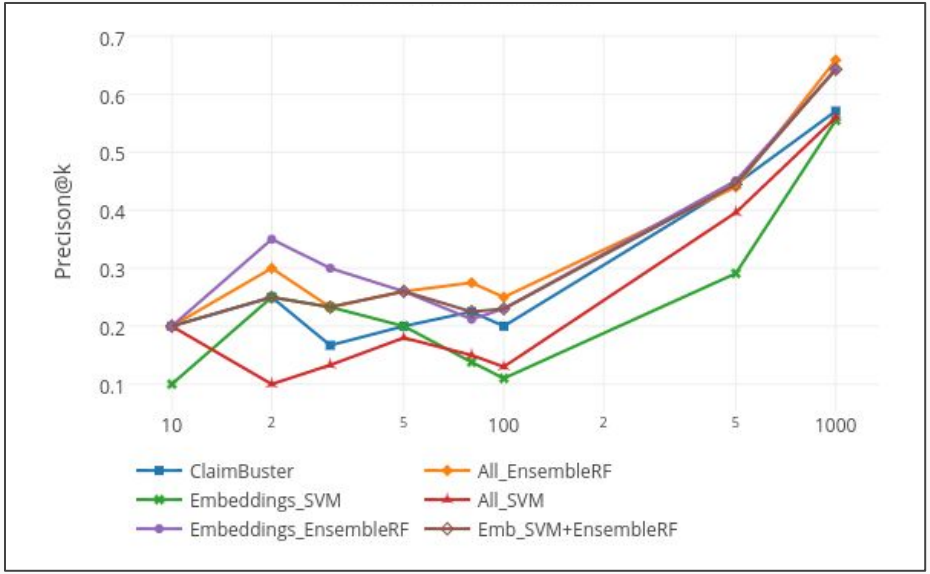
Feature Set	Classifier	Precision	Recall	F1-score
Embeddings	**SVM**	0.15	0.35	0.21
Embeddings	**EnsembleRF**	0.10	0.74	0.18
All	**SVM**	0.10	0.75	0.18
All	**EnsembleRF**	0.11	0.74	0.19

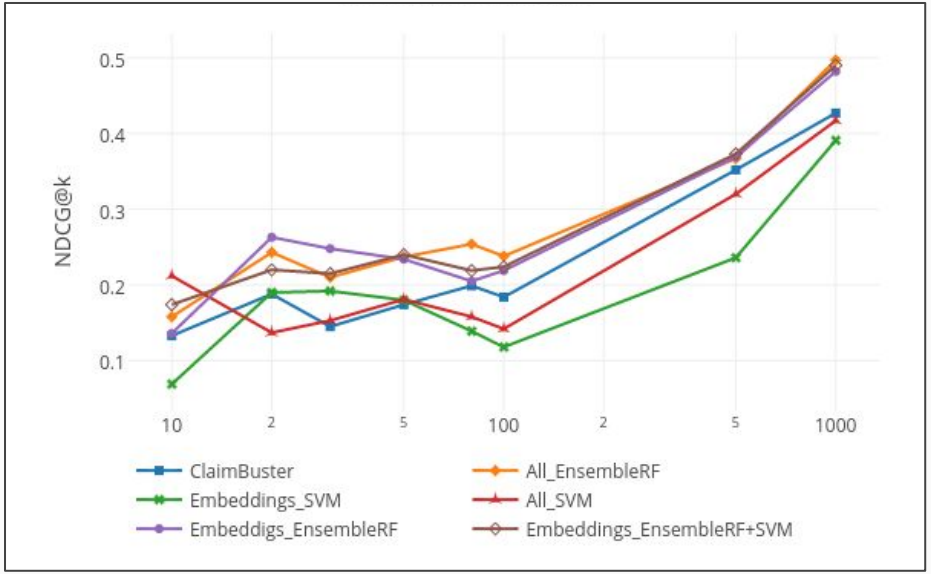
Results & Comparison: Classification

Feature Set	Classifier	Precision	Recall	F1-score
Embeddings	**SVM+EnsembleRF**	0.178	0.309	0.224
ClaimBuster	**-**	0.18	0.30	0.23

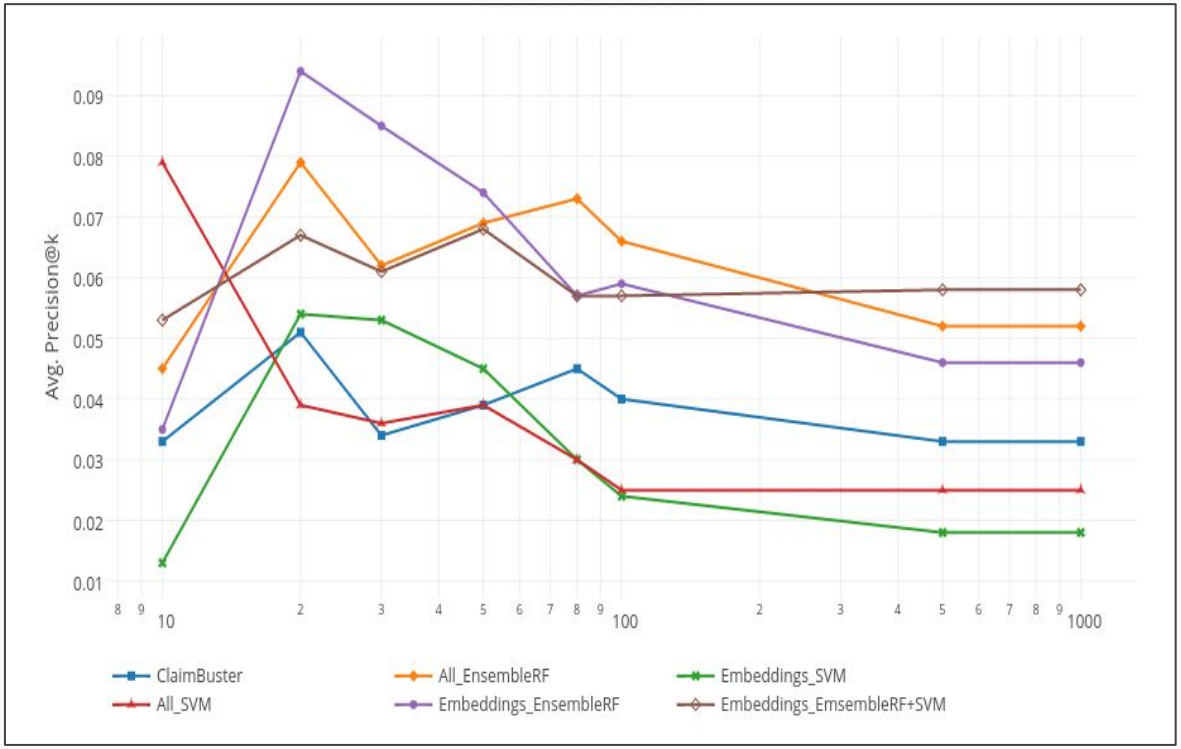
Results & Comparison: Ranking

- Scores for each example obtained using the probability of check-worthy class given by the classifier.





Results & Comparison: Ranking



Stance Detection

Information Analysis

Establishing the relationship of a claim with other facts / news articles. The relationships can be “unrelated”, “agree”, “disagree” or “discuss”.

The Dataset

- Provided by the Fake News Challenge 1.
- A headline and a body text - either from the same news article or from two different articles.
- Derived from the Emergent dataset http://www.emergent.info/

rows	unrelated	discuss	agree	disagree
49972	0.73131	0.17828	0.0736012	0.0168094

Data Partition	#Total Pairs	#Related Pairs	Percentage
Train	32400	8613	64.8%
Development	7950	2090	15.9%
Test	9622	2724	19.3%

An Example from the dataset

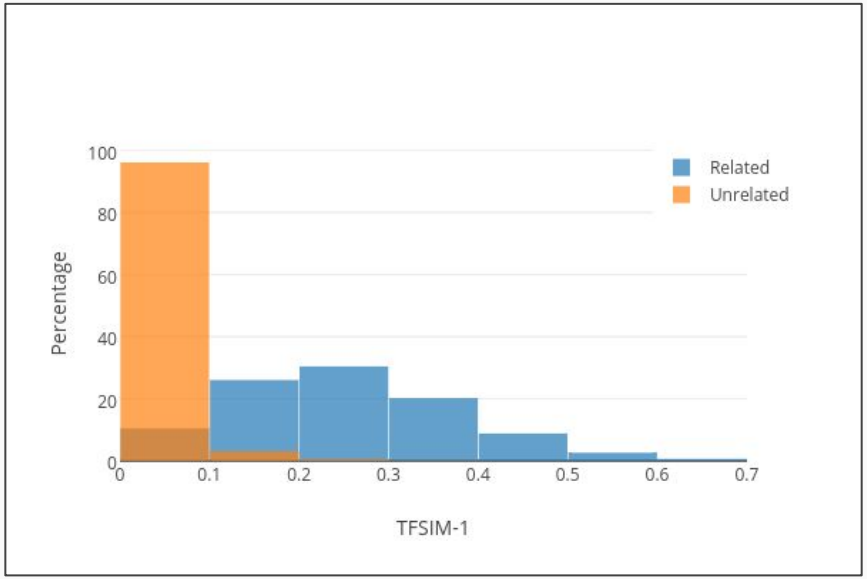
Kanye West barred from all future award shows.	Claim
...Kanye has not been banned from all future award shows...	Disagree
...After this latest incident,organizers of … have unanimously agreed to disinvite and bar West ...	Agree
“Has Kanye West been barred from all future grammy award shows?”	Discuss
“The apple watch sport may start at a mere $349.”	Unrelated

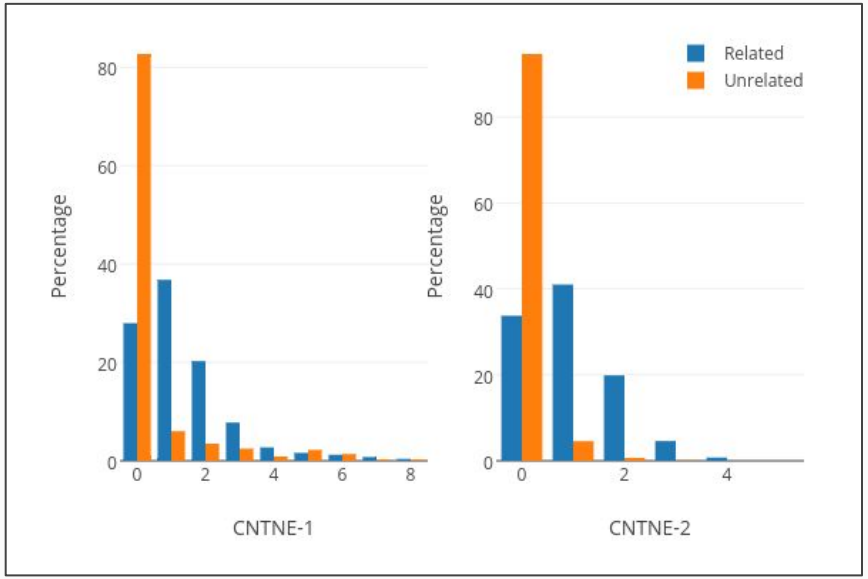
Task A

Classify into related/unrelated

Task A : Features

- TF-IDF similarity between the headline and the body text.
- Named Entity overlap. Number of NE common in headline and body. Only NE of Person, Location, Organization, Money, Percent class are considered in CNTNE-2.
- Word overlap - Number of tokens in headline present in first 100 tokens of article.





Task A : Results

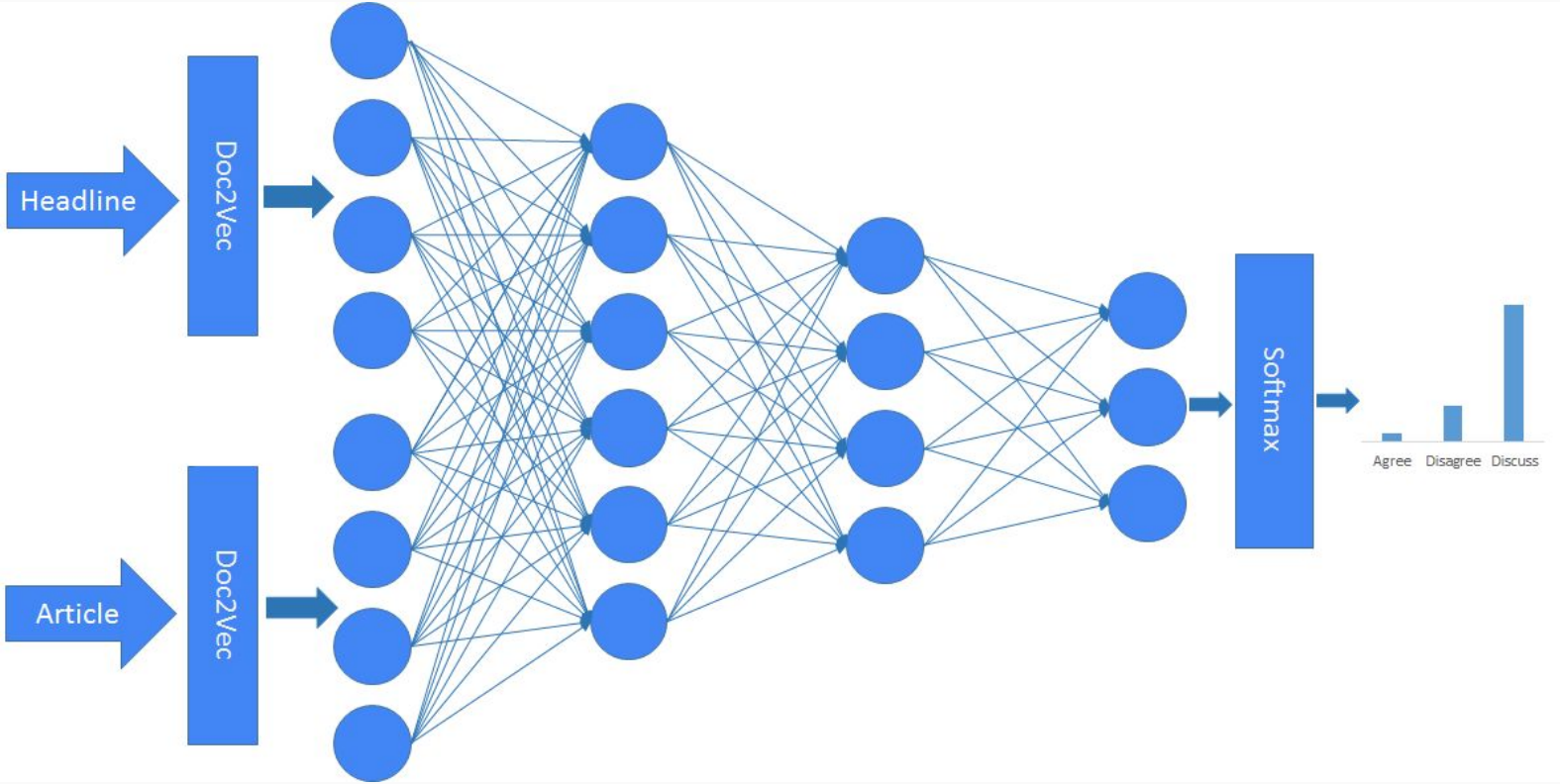
- System performs better than the official baseline.
- The metrics precision, recall and f1-score are macro-averages for the two class.

Features	Precison	Recall	F1-Score	Accuracy
TFSIM-1	0.92	0.86	0.89	91.43%
CNTNE-2	0.88	0.81	0.83	87.52%
TFSIM1+ CNTNE2	0.95	0.93	0.94	94.98%
TFSIM2+ CNTNE2	0.94	0.92	0.93	94.12%
TFSIM2+ OVLP+ CNTNE2	**0.96**	**0.95**	**0.95**	**96.35%**
Baseline OB	0.96	0.93	0.94	95.61%

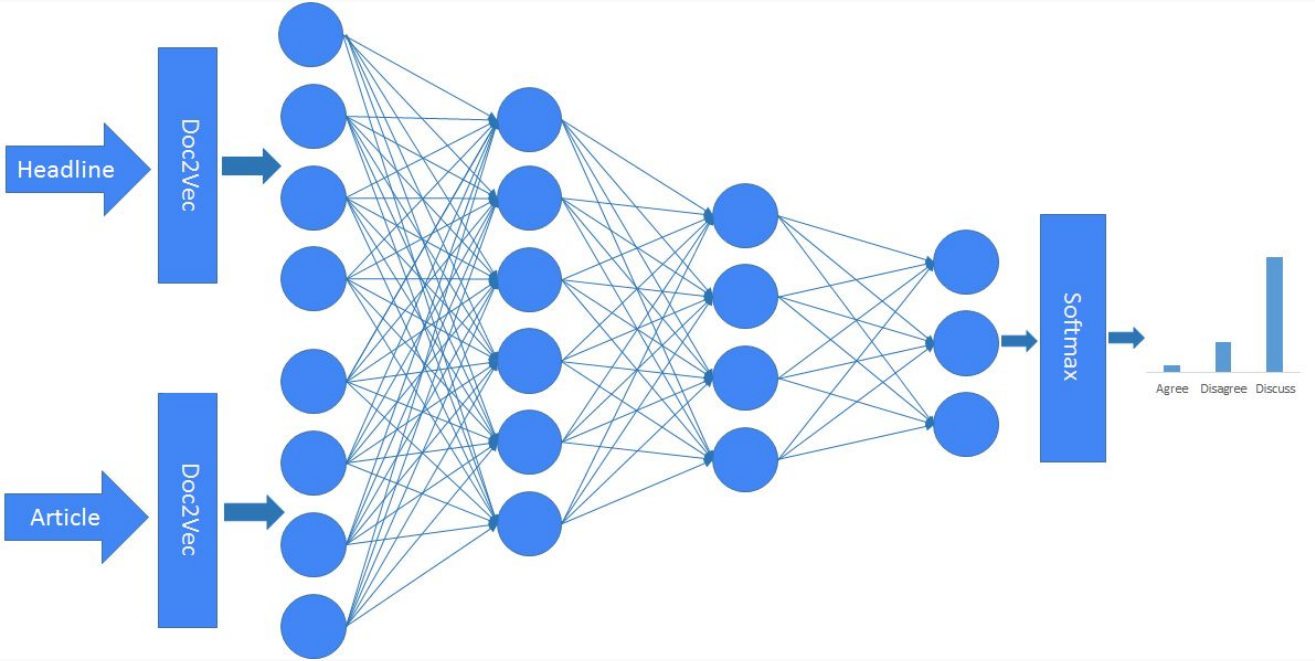
Task B

Classify related into agree/disagree/discuss.

Models: MLP with Embeddings

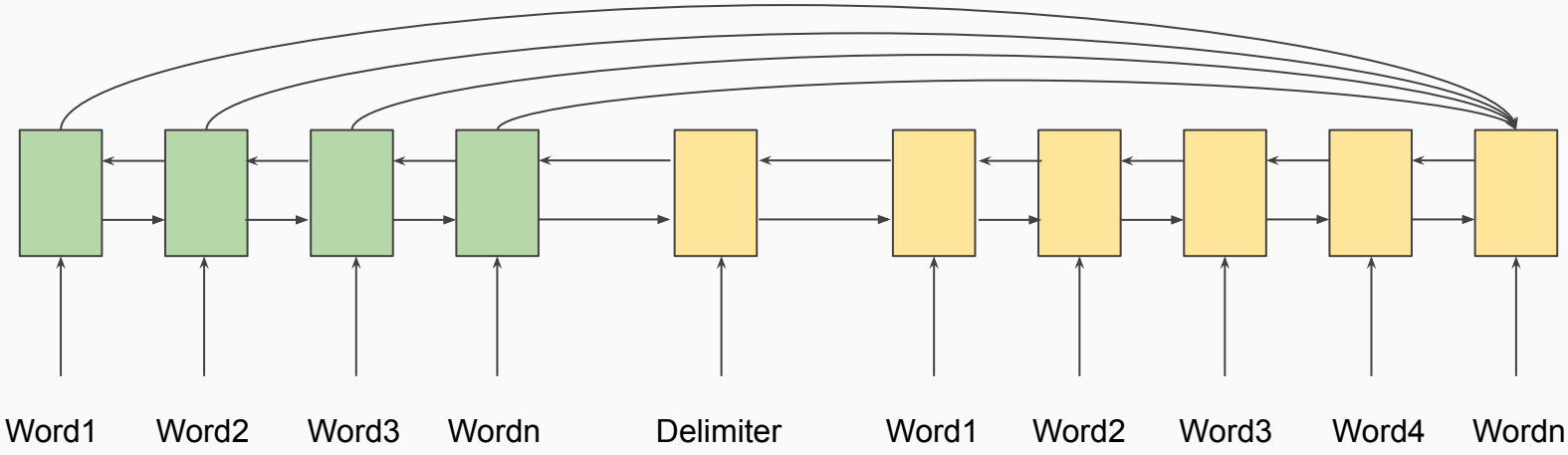


Models: MLP with Embeddings

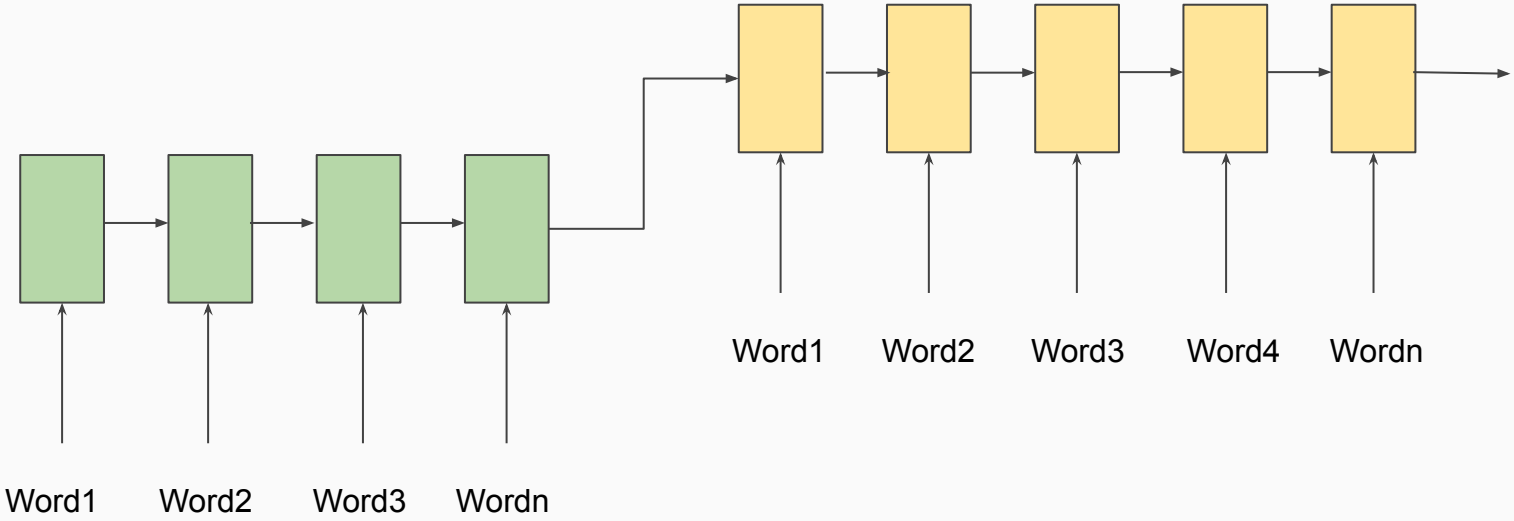


Precision	Recall	F1-score	Accuracy
0.41	0.40	0.39	68.61%

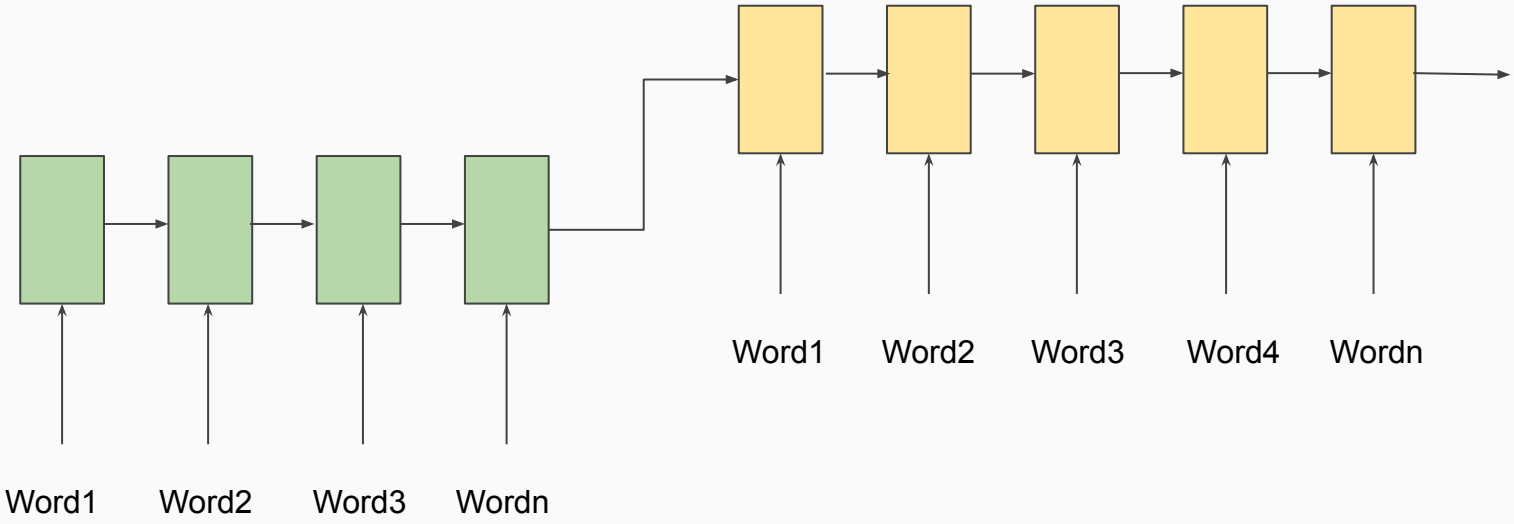
Models: Conditional Encoding



Models: Conditional Encoding

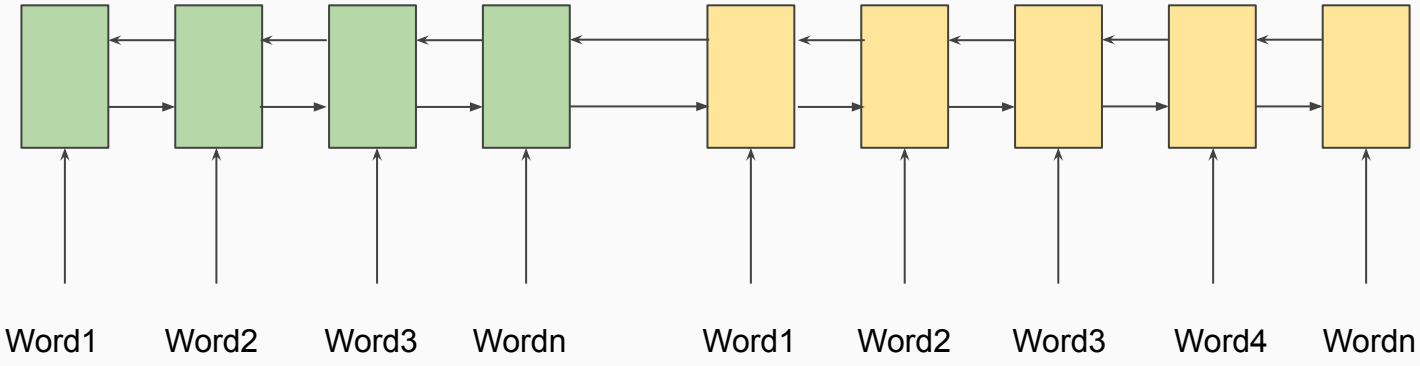


Models: Conditional Encoding

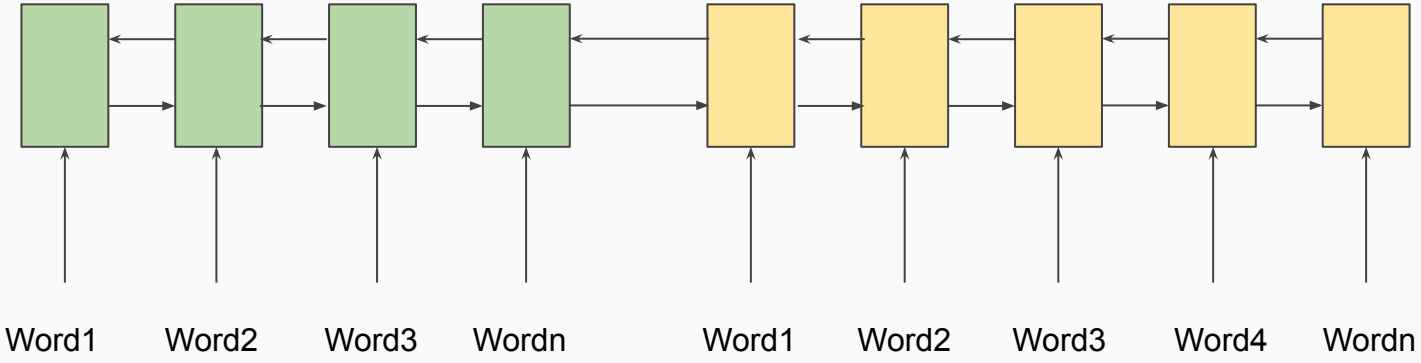


Precision	Recall	F1-score	Accuracy
0.42	0.42	0.41	68.80%

Models: BiConditional Encoding

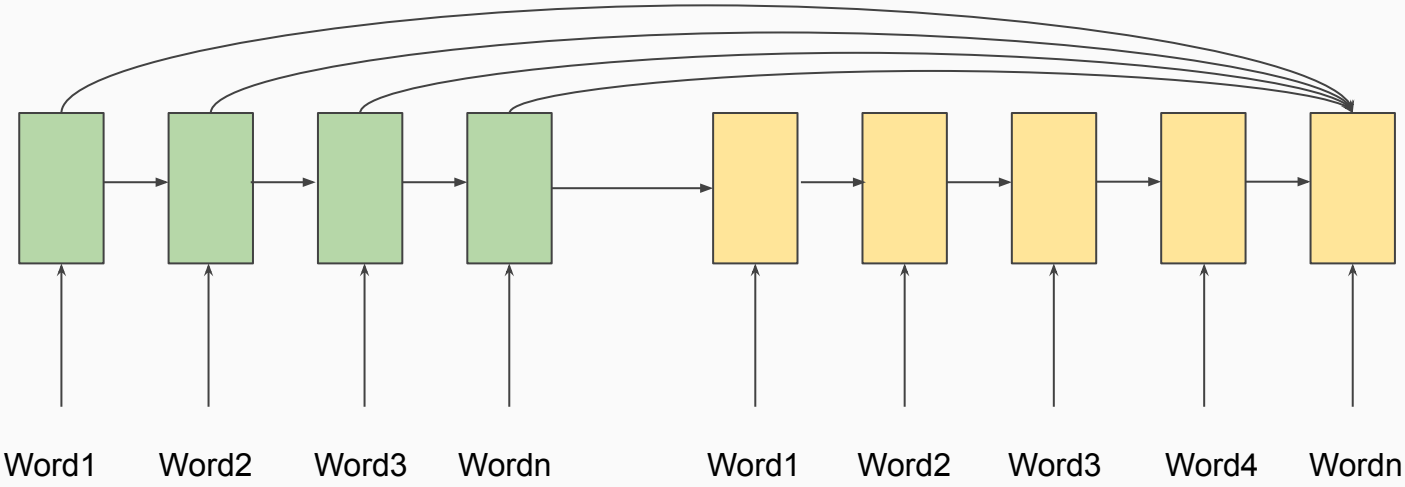


Models: BiConditional Encoding

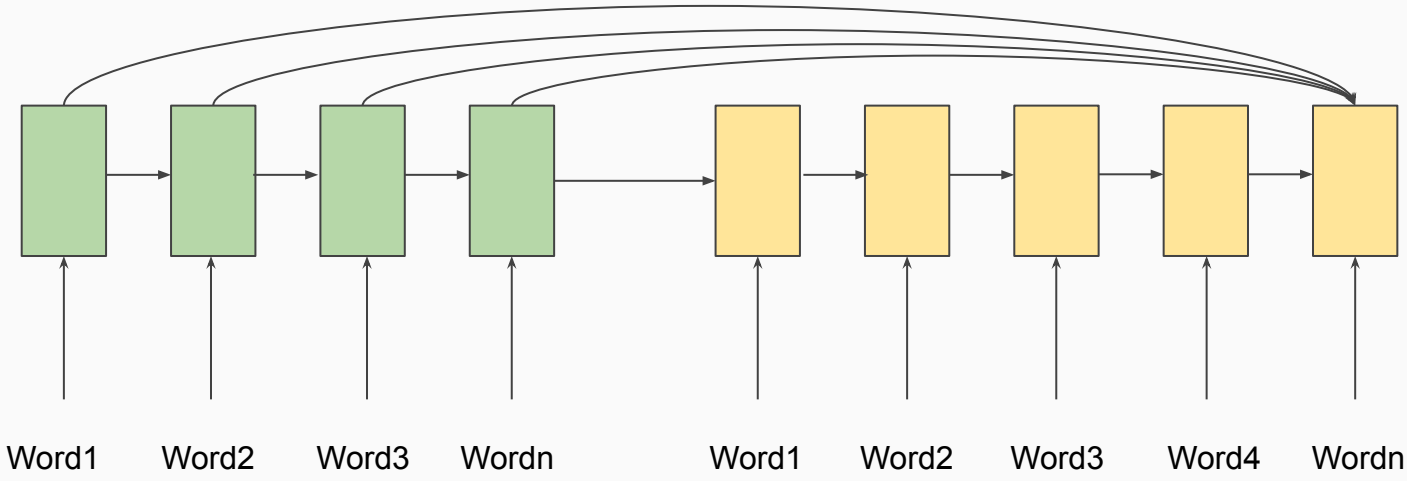


Precision	Recall	F1-score	Accuracy
0.44	0.46	0.45	71.80%

Models: Conditional Encoding With Global Attention

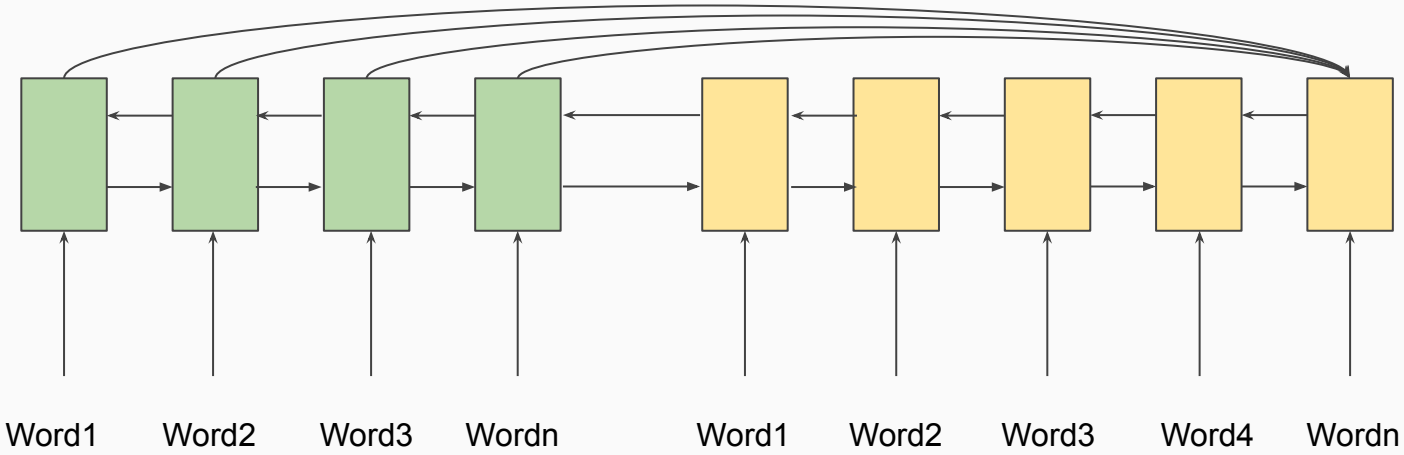


Models: Conditional Encoding With Global Attention

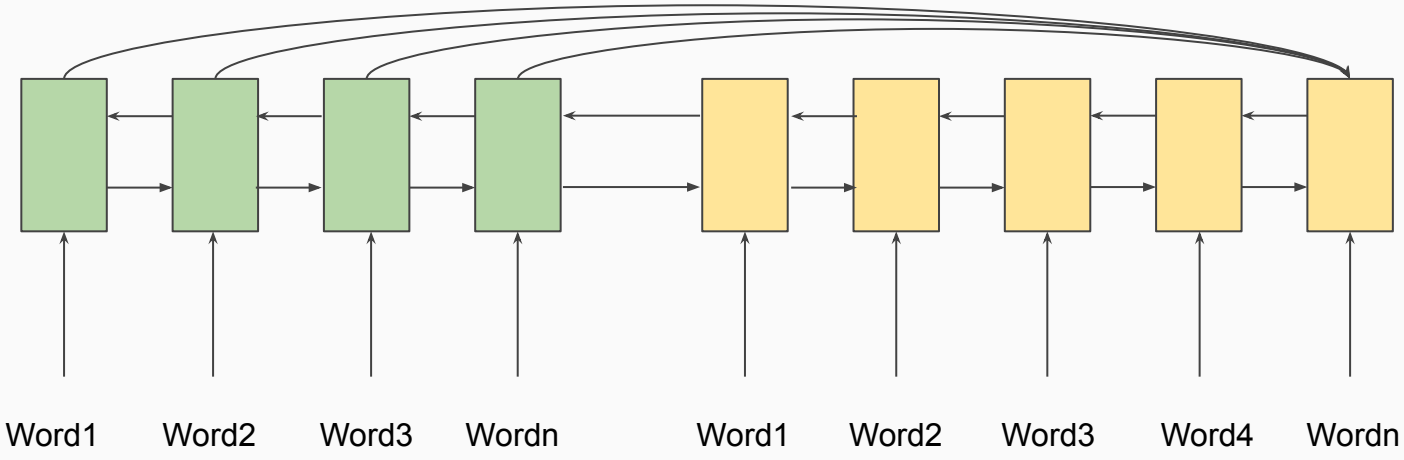


Precision	Recall	F1-score	Accuracy
0.44	0.47	0.45	70.96%

Models: BiDirectional Conditional Encoding With Global Attention

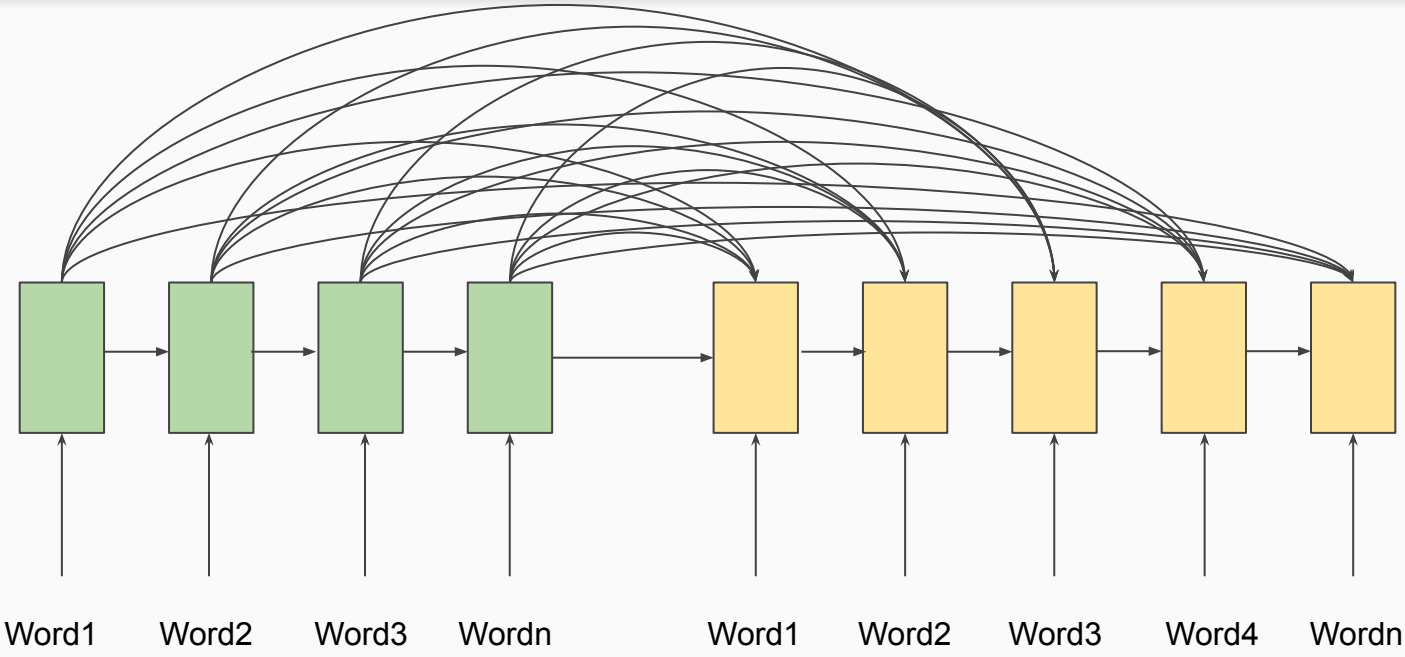


Models: BiDirectional Conditional Encoding With Global Attention



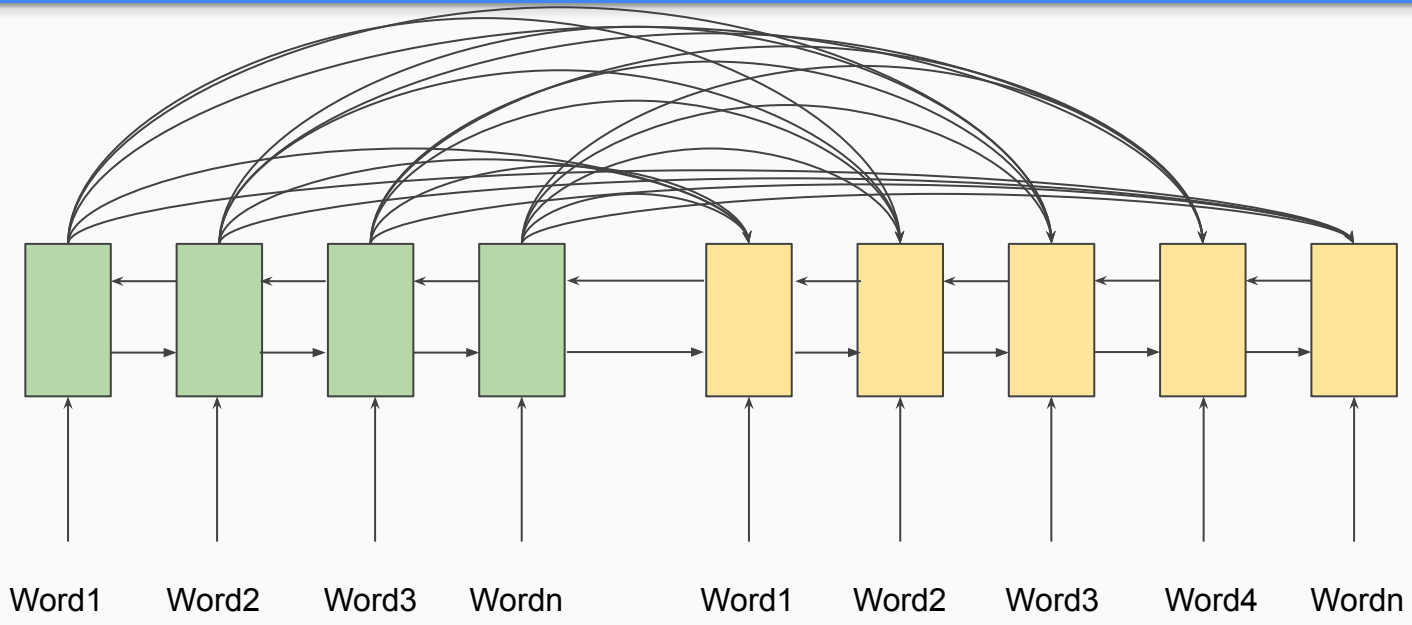
Precision	Recall	F1-score	Accuracy
0.61	0.56	0.57	74.08%

Models: Conditional Encoding With Word Attention



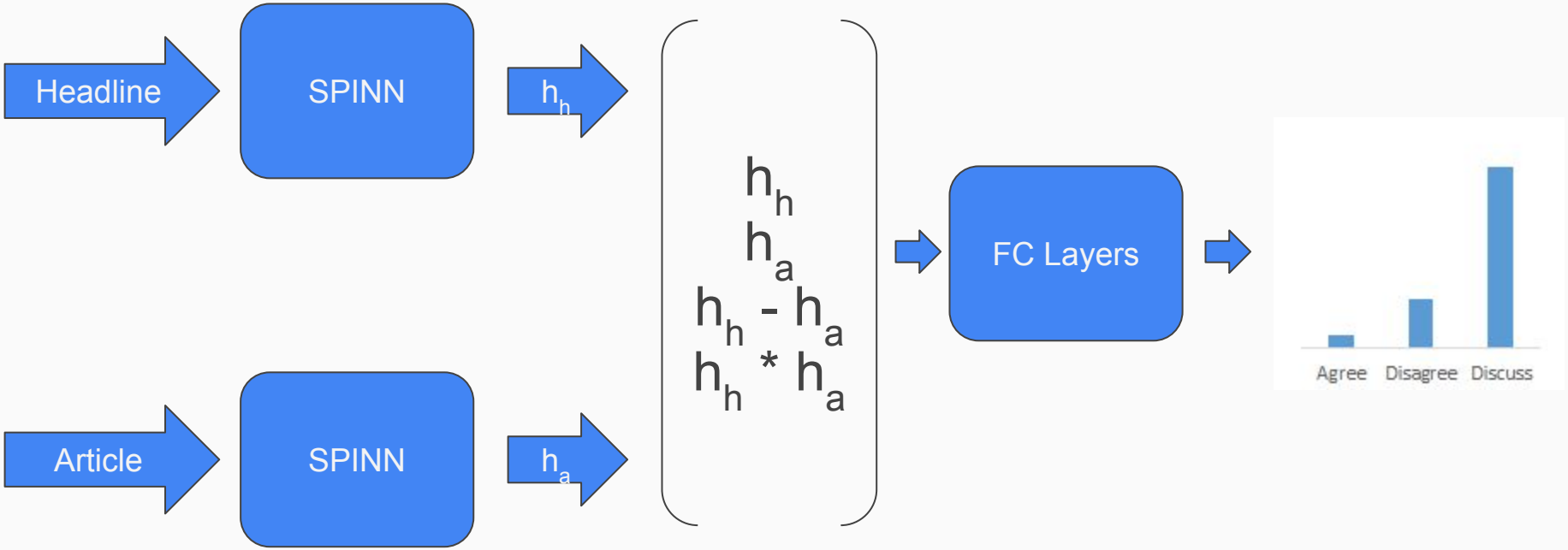
Precision	Recall	F1-score	Accuracy
0.50	0.49	0.48	70.81%

Models: BiDirectional Conditional Encoding With Word Attention

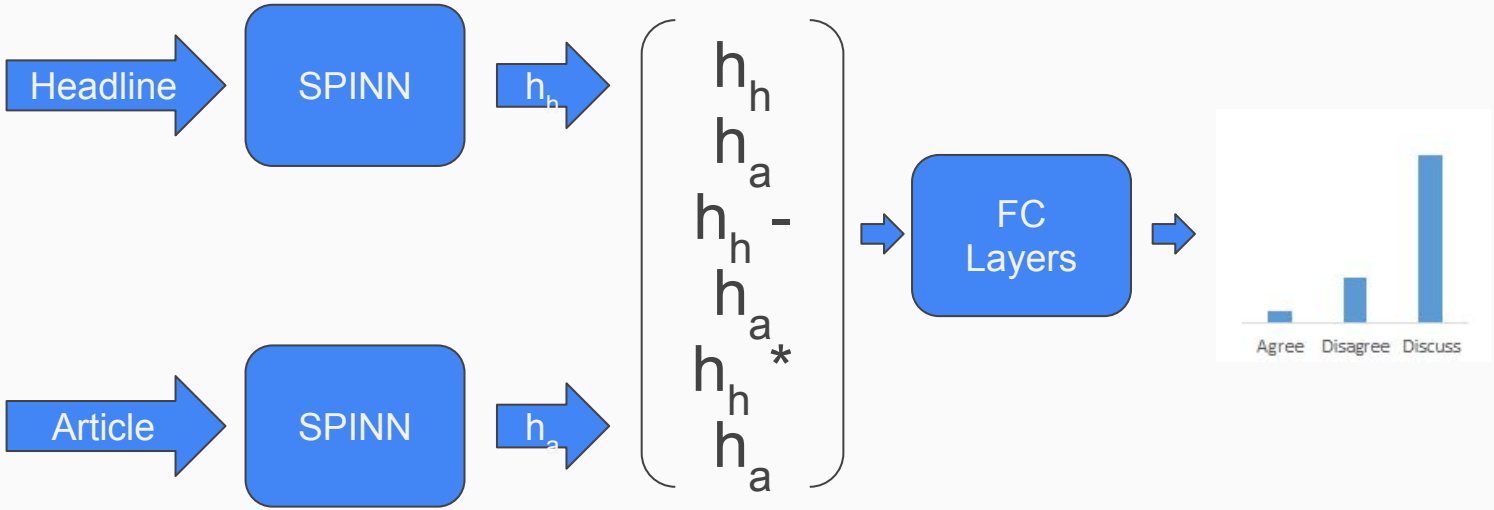


Precision	Recall	F1-score	Accuracy
0.64	0.58	0.59	74.52%

Models: Sentence Representation using SPINN



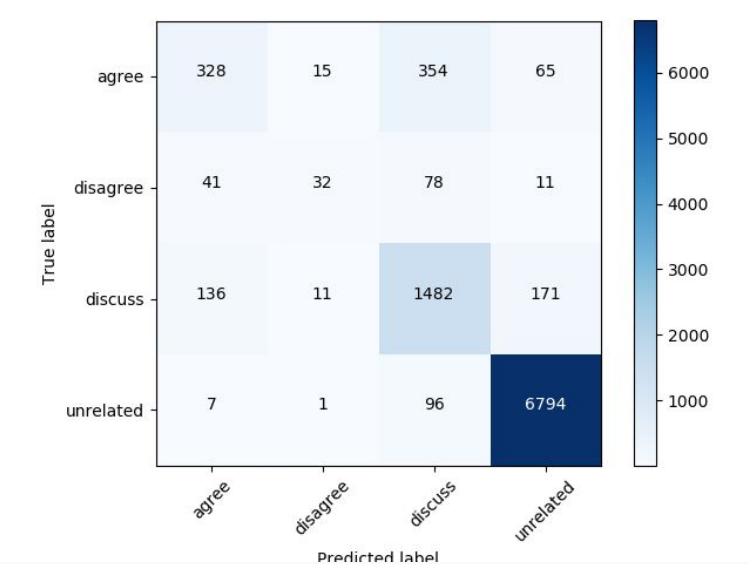
Models: Sentence Representation using SPINN

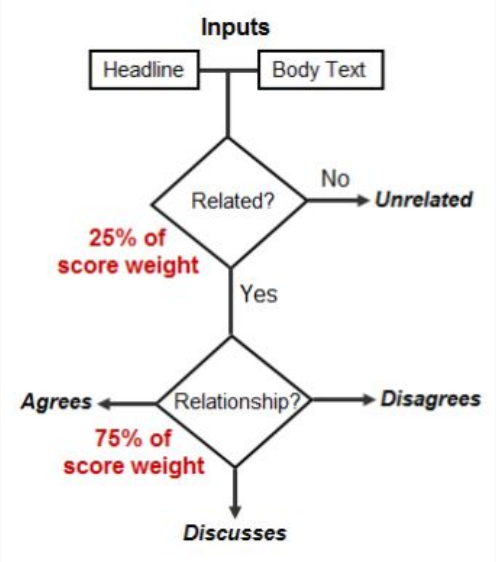


Precision	Recall	F1-score	Accuracy
0.59	0.56	0.57	74.08%

Comparison

Model	Accuracy : Task A only	Accuracy : Task B Only	FNC1 ScoreTF
Our System (BiDiWord)	96.35%	74.52%	79.60%
Official Baseline	95.61%	67.68%	79.03%





Future Work

- It should be possible to build a prototype post-facto “truth labeling” system from a “stance detection” system. Such a system would tentatively label a claim or story as true/false based on the stances taken by various news organizations on the topic, weighted by their credibility.

- We should then focus on the other stages of the pipeline : Disambiguation, converting to structural forms and identification of sources.

Thanks