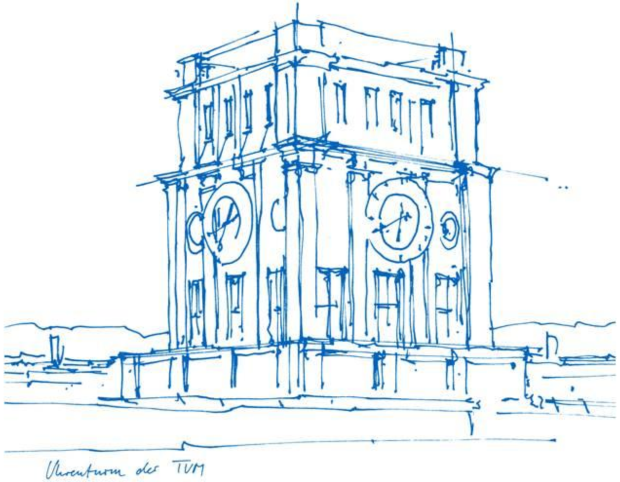
Knowledge Graphs for Better NLP Model Explanations

Technische Universität München

Fakultät für Informatik

NLP Lab Course, SS21

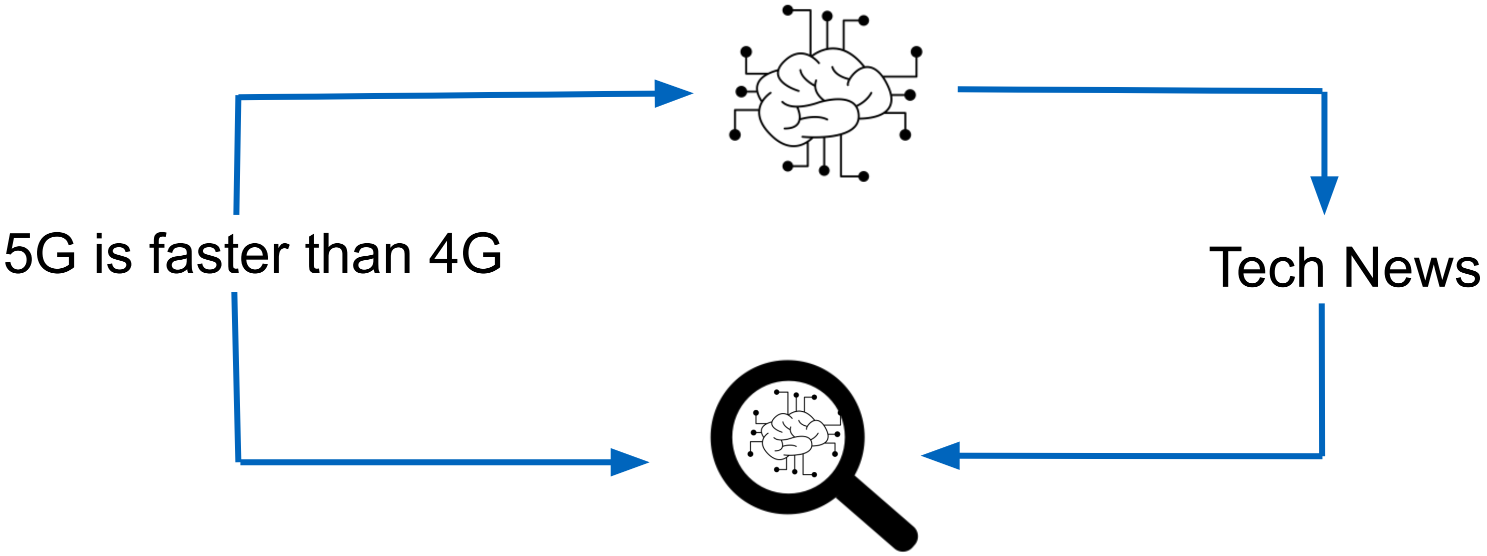
Date :16.07.2021



Hyein Koo, Anagha Moosad

Knowledge Grounded Explanations

Current State of Explainable AI

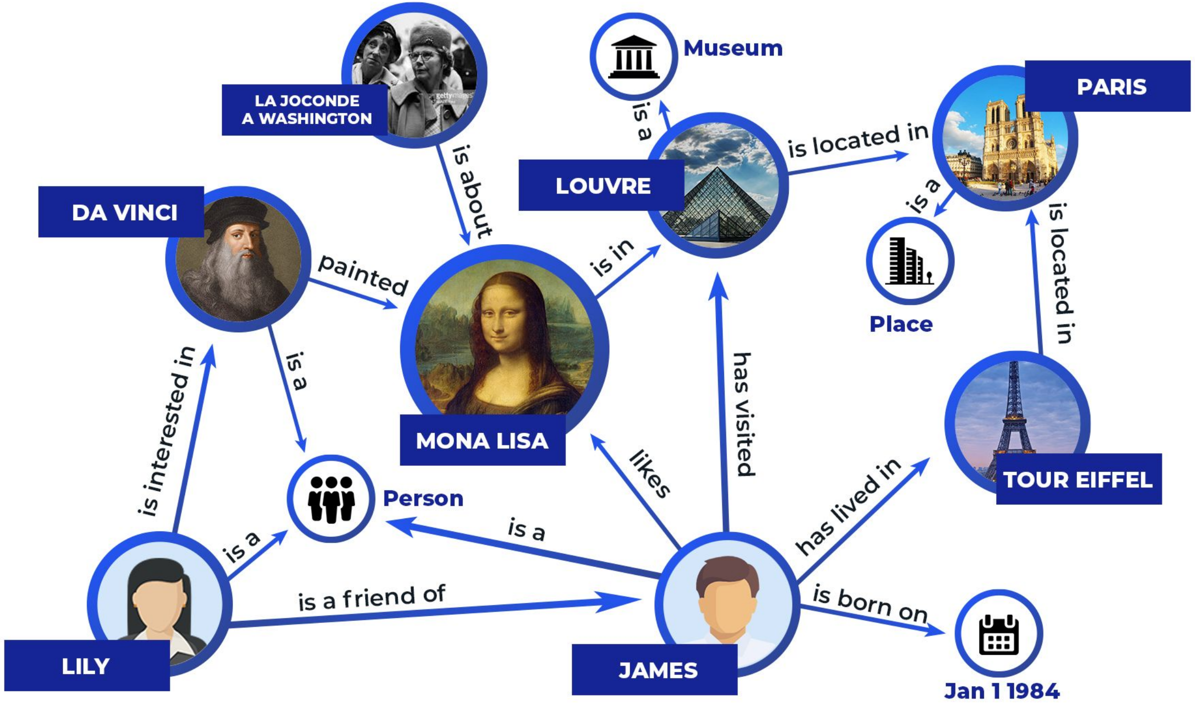


5G is faster than 4G

Tech News

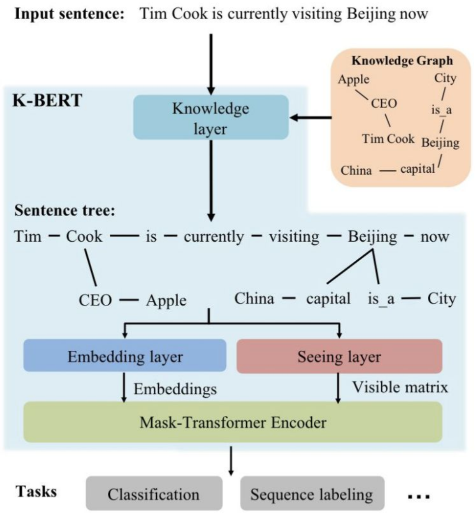
How to enrich Input for better explanations?

Enter Knowledge Graphs !!



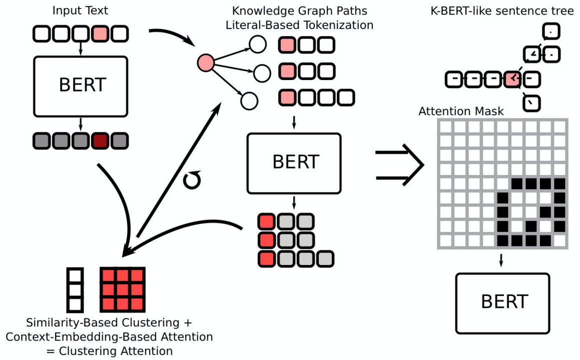
Related Work

1. K-BERT: Enabling Language Representation with Knowledge Graph



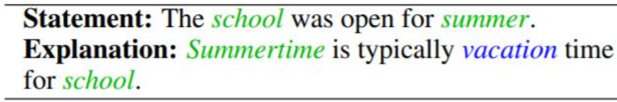
The method not tested for explainability

2. Explainable Natural Language Processing with Knowledge Graphs


Input Text
BERT
Knowledge Graph Paths
Literal-Based Tokenization
BERT
K-BERT-like sentence tree
Attention Mask
BERT
Similarity-Based Clustering +
Context-Embedding-Based Attention
= Clustering Attention

The method uses knowledge graph embeddings to get better explanations and this method is not completely implemented

3.Generating Commonsense Explanation by Extracting Bridge Concepts from Reasoning Paths



Statement: The *school* was open for *summer*.
Explanation: *Summertime* is typically *vacation* time for *school*.

The method is does not use XAI framework nor does it use explainability for text classification task.

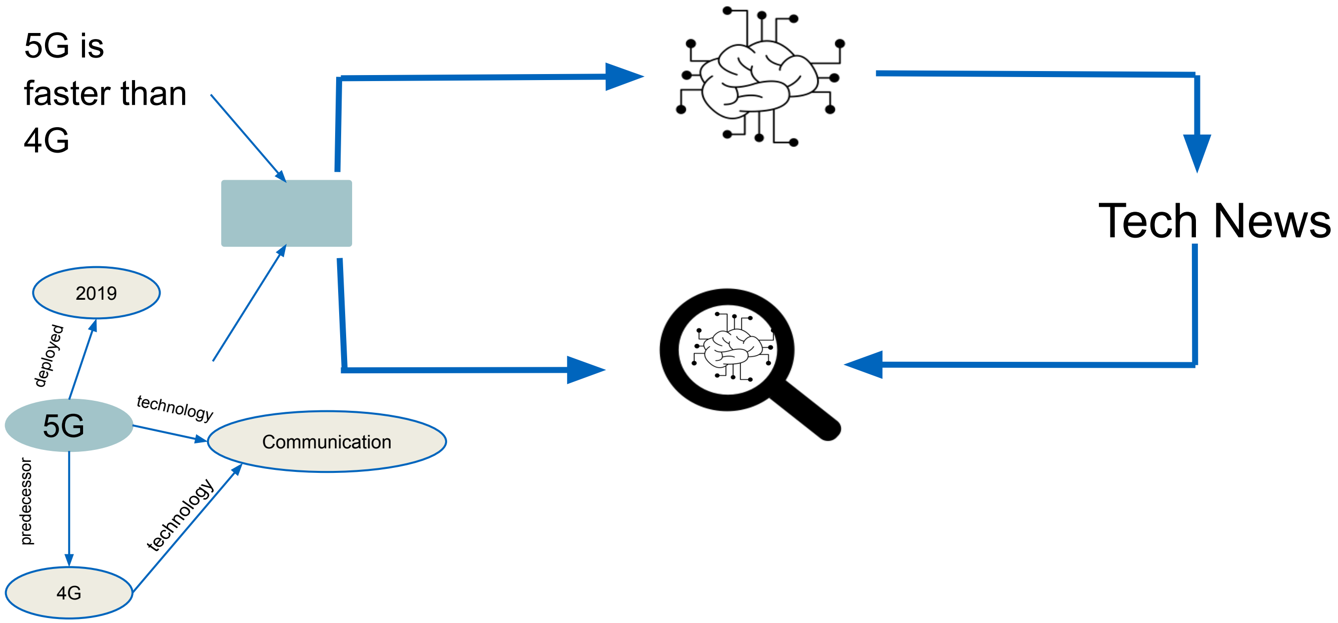
Research Question

Can we combine Input sentence and knowledge graph data to create knowledge enriched sentence as input for creating better explanations?

Can adding more than 1 hop neighbourhood data from knowledge graph create better explanations?

Project Goal

Multi-Hop K Bert Model for better explanations for text classification



Dataset

Text Data

- AGNews
- News topic classification
- 4 Classes : World, Sci-Tech, Sport, Business
- Training samples : 120,000 Testing 7,600

Knowledge Graph

- ConceptNet
- semantic network, designed to help computers understand the meanings of words that people use.

Text Data - Preprocessing

Raw Data

Class Index	Title	Description
3	Fears for T N pension after talks	Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken parent firm Federal Mogul.
4	The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (SPACE.com)	SPACE.com - TORONTO, Canada -- A second\team of rocketeers competing for the #36;10 million Ansari X Prize, a contest for\privately funded suborbital space flight, has officially announced the first\launch date for its manned rocket.

Preprocessed Data

Class Index	Text
2	unions representing workers turner newall say disappointed talks stricken parent firm federal mogul
3	spacecom toronto canada secondteam rocketeers competing three thousand, six hundred and ten million ansari x prize contest forprivately funded suborbital space flight officially announced firstlaunch date manned rocket

Knowledge Graph - Preprocessing

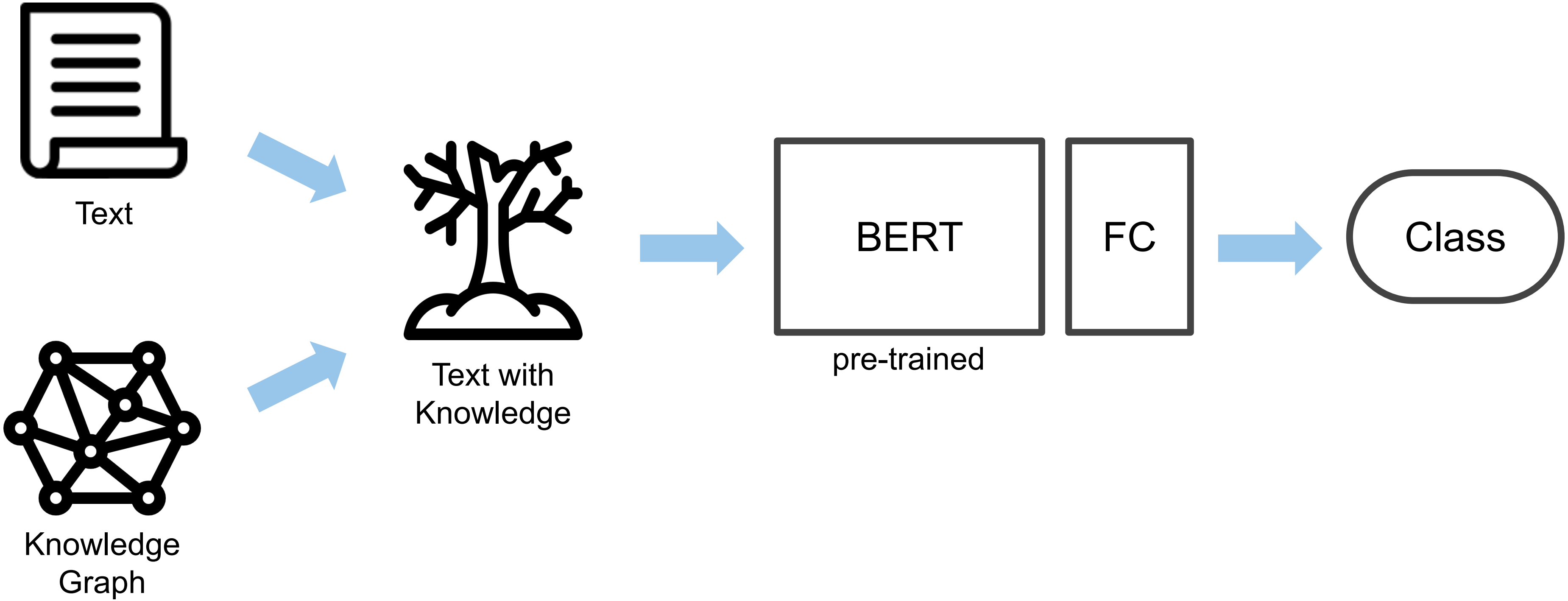
Raw Data

URI	Relation	Term	Label
/a/[/r/Causes/,/c/en/relaxing/,/c/en/increased_mental_clarity/]	/r/Causes	/c/en/relaxing	/c/en/increased_mental_clarity
/a/[/r/CreatedBy/,/c/en/software/,/c/en/programmer/]	/r/CreatedBy	/c/en/software	/c/en/programmer

Preprocessed Data

Subject	Relation	Object
relaxing	causes	inattention_to_detail
software	created_by	programmer

Model Architecture



Sentence Tree

As Long As A Basketball Court: Australia's Largest Dinosaur Confirmed

Lowercase, multi-word tokenization

as	long	as	a	basketball_court	:	australia	‘	s	largest	dinosaur	confirmed

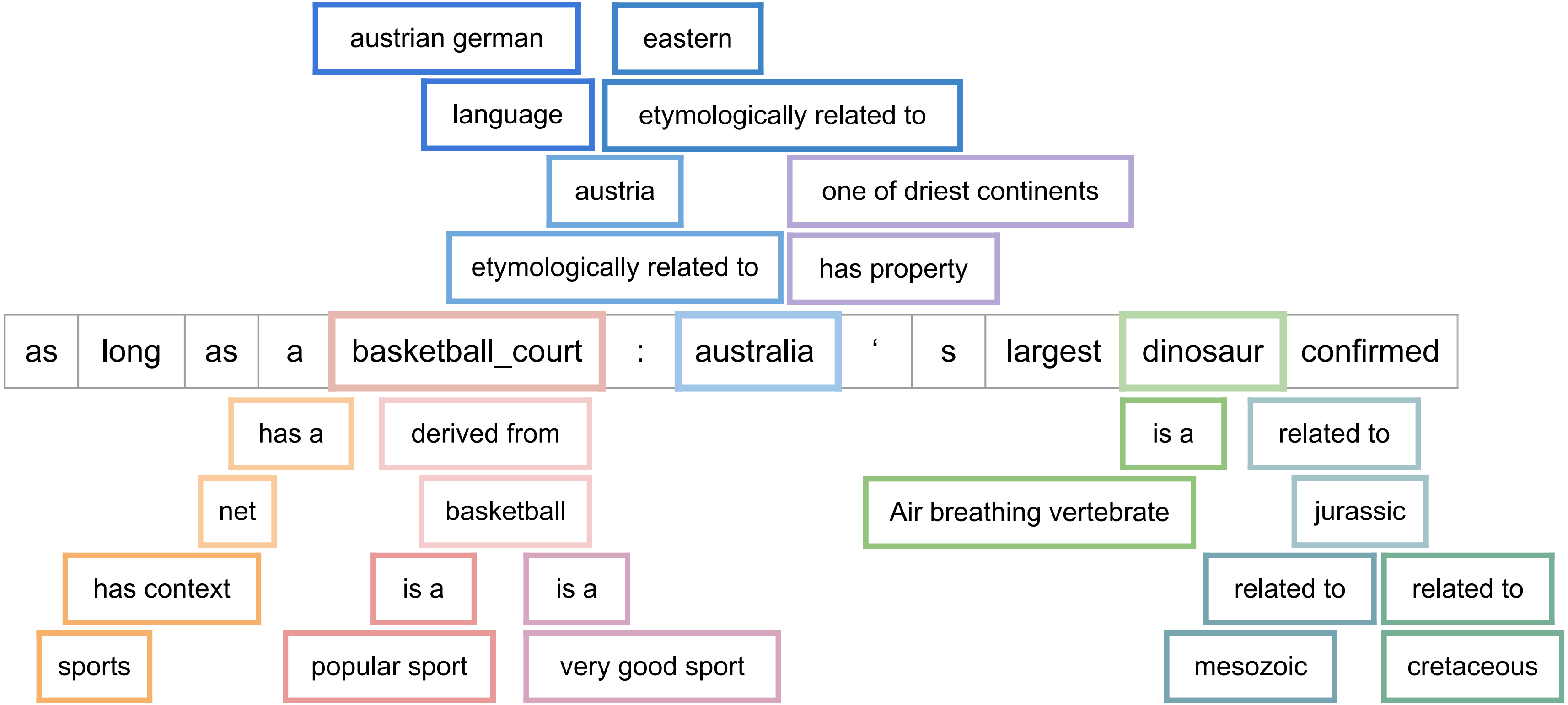
POS tagging

as	long	as	a	basketball_court	:	australia	‘	s	largest	dinosaur	confirmed
RB	RB	IN	DT	NN	:	NN	‘’	RB	JJS	NN	VBD

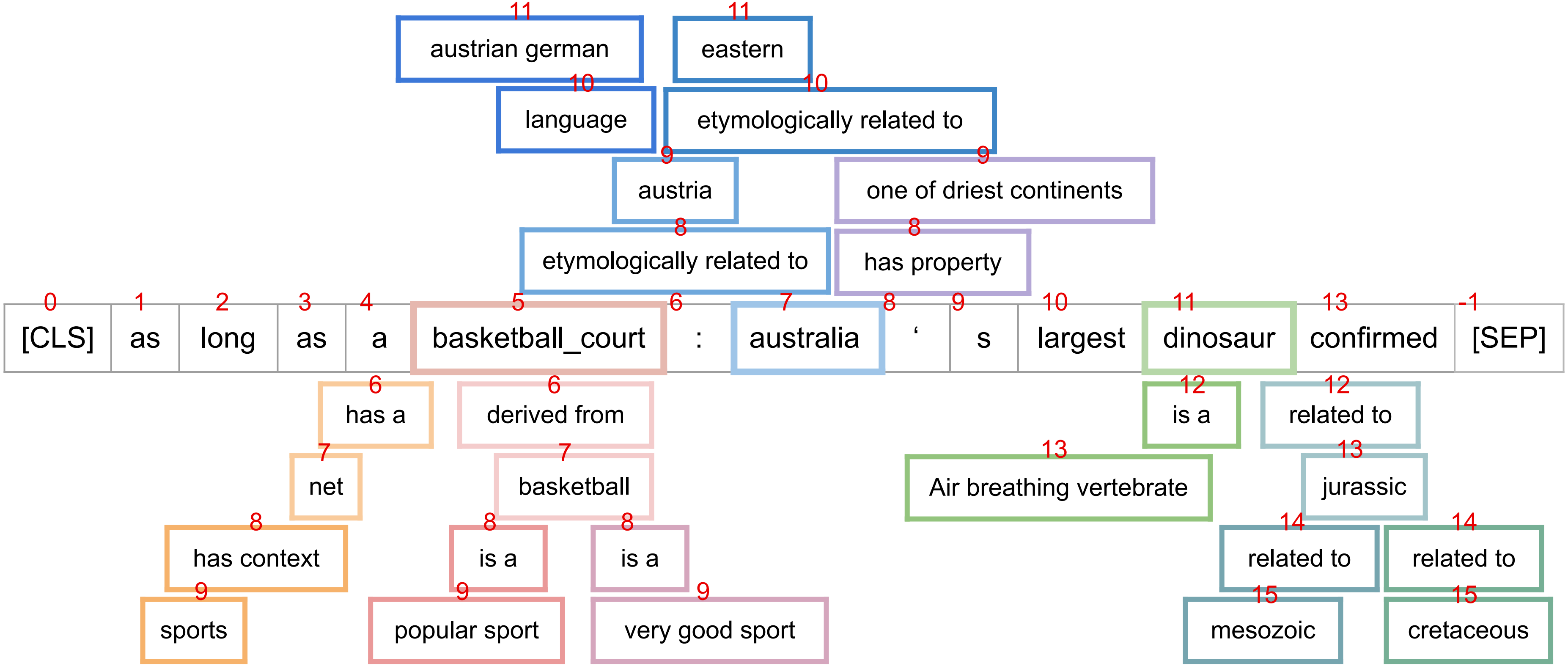
Relations

Antonym	Entails	Has property	Not has property	dbpedia/capital
At location	Etymologically derived from	Has subevent	Not used for	dbpedia/field
Capable of	Etymologically related to	Instance of	Obstructed by	dbpedia/genre
Causes	External URL	Is a	Part of	dbpedia/genus
Causes desire	Form of	Located near	Receives action	dbpedia/influenced by
Created by	Has a	Made of	Related to	dbpedia/known for
Defined as	Has context	Manner of	Similar to	dbpedia/language
Derived from	Has first subevent	Motivated by goal	Symbol of	dbpedia/leader
Desires	Has last subevent	Not capable of	synonym	dbpedia/occupation
Distinct from	Has prerequisite	Not desires	Used for	dbpedia/product

Sentence Tree



Sentence Tree



0	1	2	3	4	5	6	7	8	9	10	11	13	-1
[CLS]	as	long	as	a	basketball_court	:	australia	‘	s	largest	dinosaur	confirmed	[SEP]

11 austrian german

11 eastern

10 language

10 etymologically related to

9 austria

9 one of driest continents

8 etymologically related to

8 has property

6 has a

6 derived from

12 is a

12 related to

7 net

7 basketball

13 Air breathing vertebrate

13 jurassic

8 has context

8 is a

8 is a

14 related to

14 related to

9 sports

9 popular sport

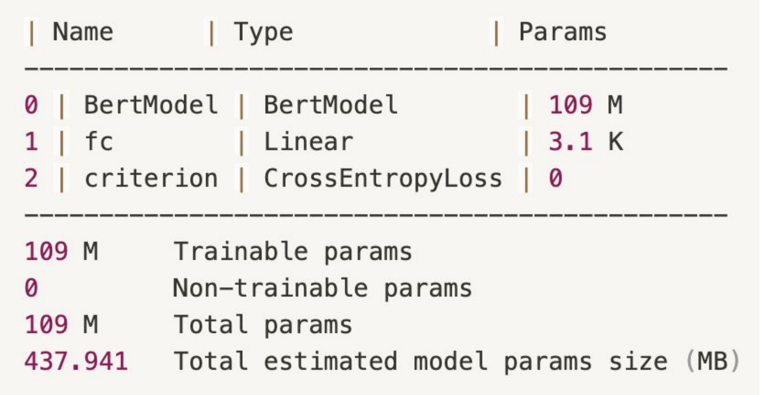
9 very good sport

15 mesozoic

15 cretaceous

Training Setting

- Dataset: 96000(train), 24000(val), 7600(test)
- Max epoch: 10
- Adam Optimizer
- Learning rate: 0.00001

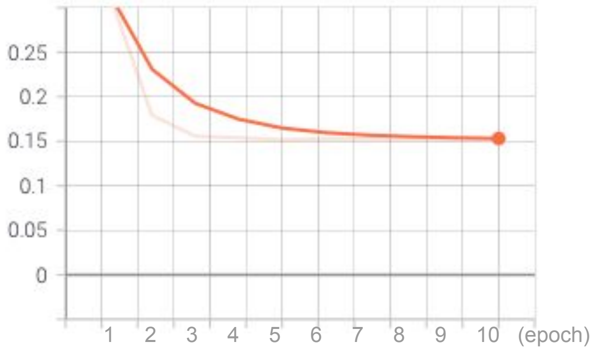


```
  | Name      | Type             | Params
-----------------------------------------------
0 | BertModel | BertModel        | 109 M
1 | fc        | Linear           | 3.1 K
2 | criterion | CrossEntropyLoss | 0
-----------------------------------------------
109 M     Trainable params
0         Non-trainable params
109 M     Total params
437.941   Total estimated model params size (MB)
```

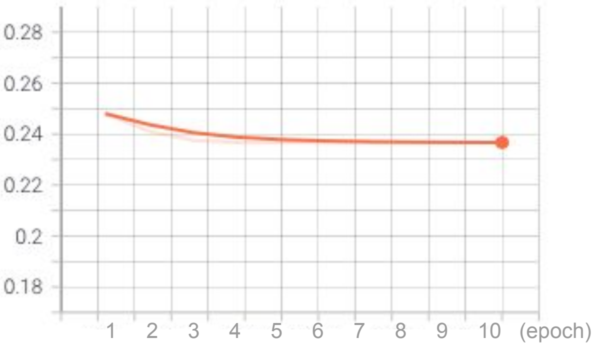
Performance

	Multi-hop K-Bert
Accuracy	90.21 %

train_loss_epoch

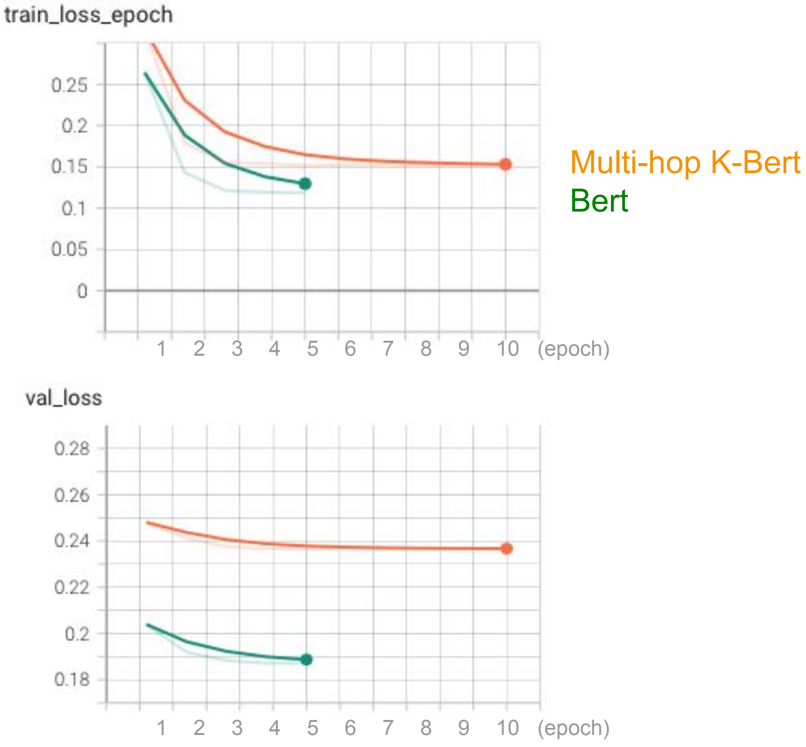


val_loss



Performance

	Multi-hop K-Bert	Bert
Accuracy	90.21 %	91.91%
Training Time	2d 22h 23m (10 epochs)	11h 18m (5 epochs)



Explainability

Captum

Primary Attribution

evaluates contribution of each input feature to the output of the model

Layer Attribution

evaluates contribution of each neuron in a given layer to the output of the model

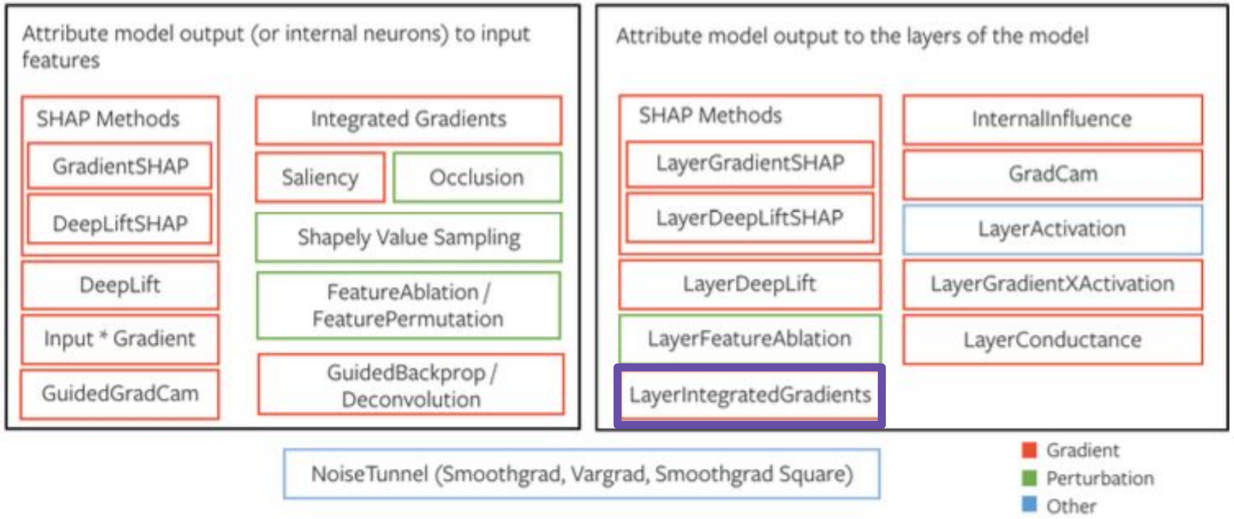
Neuron Attribution

evaluates contribution of each input feature on the activation of a particular hidden neuron

Explainability

Primary / neuron attribution algorithms

Layer attribution algorithms



Attribute model output (or internal neurons) to input features

SHAP Methods

GradientSHAP

DeepLiftSHAP

DeepLift

Input * Gradient

GuidedGradCam

Integrated Gradients

Saliency

Occlusion

Shapely Value Sampling

FeatureAblation / FeaturePermutation

GuidedBackprop / Deconvolution

Attribute model output to the layers of the model

SHAP Methods

LayerGradientSHAP

LayerDeepLiftSHAP

LayerDeepLift

LayerFeatureAblation

LayerIntegratedGradients

InternalInfluence

GradCam

LayerActivation

LayerGradientXActivation

LayerConductance

NoiseTunnel (Smoothgrad, Vargrad, Smoothgrad Square)

🟥 Gradient

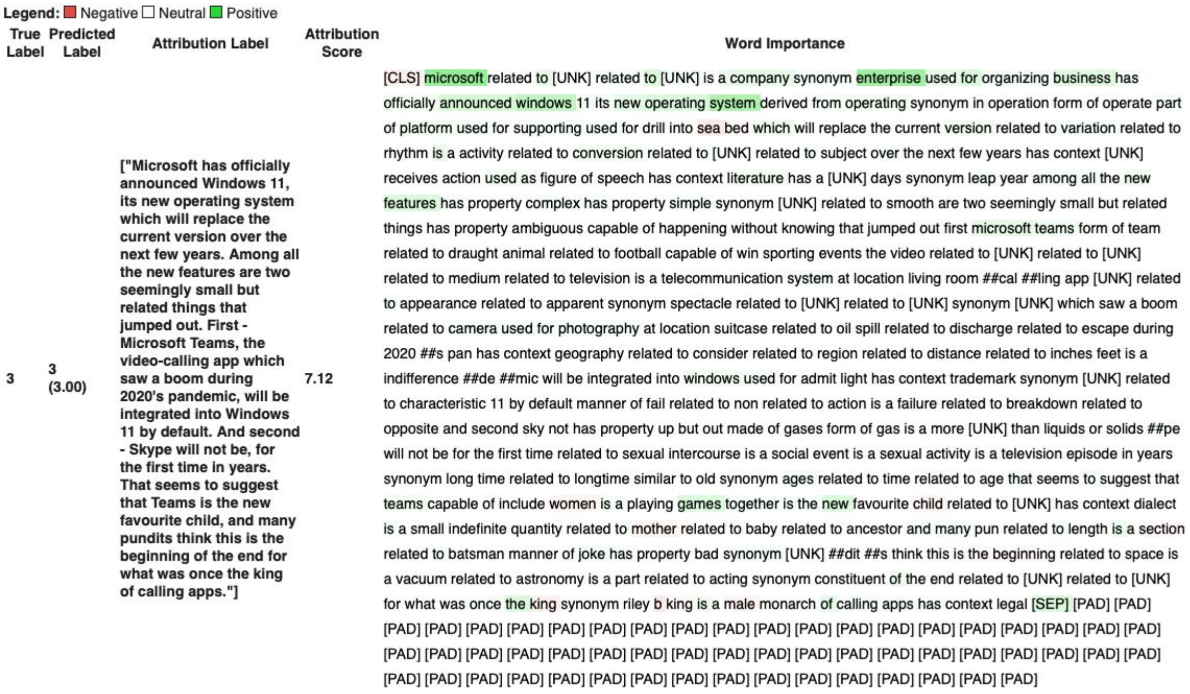
🟩 Perturbation

🟦 Other

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

Results

Captum Visualization (sci-tech news)

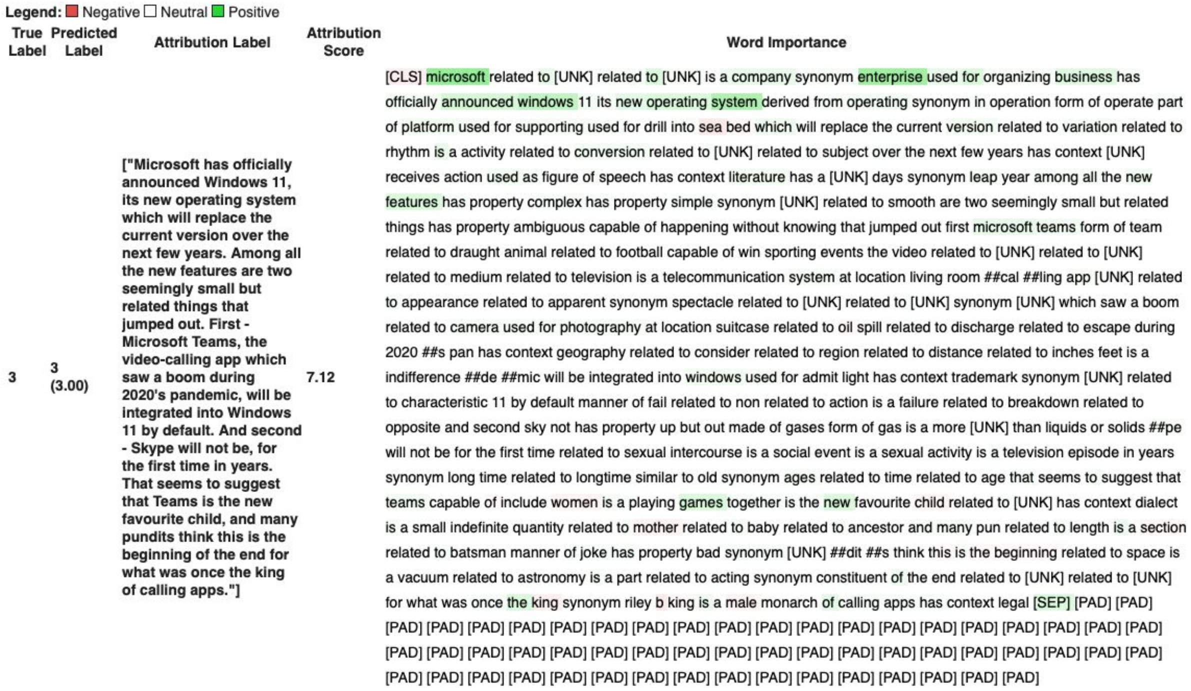


Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
3	3 (3.00)	["Microsoft has officially announced Windows 11, its new operating system which will replace the current version over the next few years. Among all the new features are two seemingly small but related things that jumped out. First - Microsoft Teams, the video-calling app which saw a boom during 2020's pandemic, will be integrated into Windows 11 by default. And second - Skype will not be, for the first time in years. That seems to suggest that Teams is the new favourite child, and many pundits think this is the beginning of the end for what was once the king of calling apps."]	7.12	[CLS] microsoft related to [UNK] related to [UNK] is a company synonym enterprise used for organizing business has officially announced windows 11 its new operating system derived from operating synonym in operation form of operate part of platform used for supporting used for drill into sea bed which will replace the current version related to variation related to rhythm is a activity related to conversion related to [UNK] related to subject over the next few years has context [UNK] receives action used as figure of speech has context literature has a [UNK] days synonym leap year among all the new features has property complex has property simple synonym [UNK] related to smooth are two seemingly small but related things has property ambiguous capable of happening without knowing that jumped out first microsoft teams form of team related to draught animal related to football capable of win sporting events the video related to [UNK] related to [UNK] related to medium related to television is a telecommunication system at location living room ##cal ##ling app [UNK] related to appearance related to apparent synonym spectacle related to [UNK] related to [UNK] synonym [UNK] which saw a boom related to camera used for photography at location suitcase related to oil spill related to discharge related to escape during 2020 ##s pan has context geography related to consider related to region related to distance related to inches feet is a indifference ##de ##mic will be integrated into windows used for admit light has context trademark synonym [UNK] related to characteristic 11 by default manner of fail related to non related to action is a failure related to breakdown related to opposite and second sky not has property up but out made of gases form of gas is a more [UNK] than liquids or solids ##pe will not be for the first time related to sexual intercourse is a social event is a sexual activity is a television episode in years synonym long time related to longtime similar to old synonym ages related to time related to age that seems to suggest that teams capable of include women is a playing games together is the new favourite child related to [UNK] has context dialect is a small indefinite quantity related to mother related to baby related to ancestor and many pun related to length is a section related to batsman manner of joke has property bad synonym [UNK] ##dit ##s think this is the beginning related to space is a vacuum related to astronomy is a part related to acting synonym constituent of the end related to [UNK] related to [UNK] for what was once the king synonym riley b king is a male monarch of calling apps has context legal [SEP] [PAD]

Results

Captum Visualization (sci-tech news)



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
3	3 (3.00)	["Microsoft has officially announced Windows 11, its new operating system which will replace the current version over the next few years. Among all the new features are two seemingly small but related things that jumped out. First - Microsoft Teams, the video-calling app which saw a boom during 2020's pandemic, will be integrated into Windows 11 by default. And second - Skype will not be, for the first time in years. That seems to suggest that Teams is the new favourite child, and many pundits think this is the beginning of the end for what was once the king of calling apps."]	7.12	[CLS] microsoft related to [UNK] related to [UNK] is a company synonym enterprise used for organizing business has officially announced windows 11 its new operating system derived from operating synonym in operation form of operate part of platform used for supporting used for drill into sea bed which will replace the current version related to variation related to rhythm is a activity related to conversion related to [UNK] related to subject over the next few years has context [UNK] receives action used as figure of speech has context literature has a [UNK] days synonym leap year among all the new features has property complex has property simple synonym [UNK] related to smooth are two seemingly small but related things has property ambiguous capable of happening without knowing that jumped out first microsoft teams form of team related to draught animal related to football capable of win sporting events the video related to [UNK] related to [UNK] related to medium related to television is a telecommunication system at location living room ##cal ##ling app [UNK] related to appearance related to apparent synonym spectacle related to [UNK] related to [UNK] synonym [UNK] which saw a boom related to camera used for photography at location suitcase related to oil spill related to discharge related to escape during 2020 ##s pan has context geography related to consider related to region related to distance related to inches feet is a indifference ##de ##mic will be integrated into windows used for admit light has context trademark synonym [UNK] related to characteristic 11 by default manner of fail related to non related to action is a failure related to breakdown related to opposite and second sky not has property up but out made of gases form of gas is a more [UNK] than liquids or solids ##pe will not be for the first time related to sexual intercourse is a social event is a sexual activity is a television episode in years synonym long time related to longtime similar to old synonym ages related to time related to age that seems to suggest that teams capable of include women is a playing games together is the new favourite child related to [UNK] has context dialect is a small indefinite quantity related to mother related to baby related to ancestor and many pun related to length is a section related to batsman manner of joke has property bad synonym [UNK] ##dit ##s think this is the beginning related to space is a vacuum related to astronomy is a part related to acting synonym constituent of the end related to [UNK] related to [UNK] for what was once the king synonym riley b king is a male monarch of calling apps has context legal [SEP] [PAD]

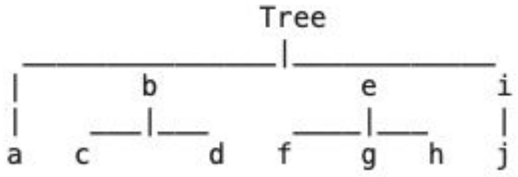
Microsoft, announced, company, enterprise, windows, new, operating, system, teams, used, for, business, new

sea, women, child, mother, king, male

Visualization

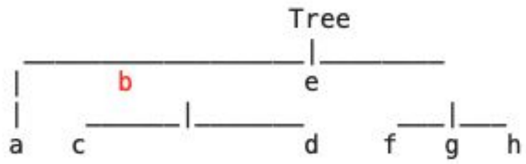
nltk.tree

- class for representing hierarchical language structures, such as syntax trees and morphological trees

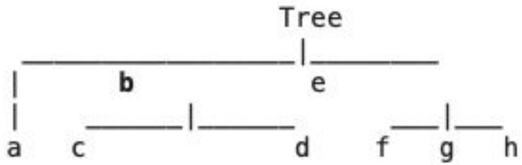


```
                 Tree
  _________________|______________
 |         b              e       i
 |      ___|___       ____|___    |
 a     c       d     f    g    h  j
```

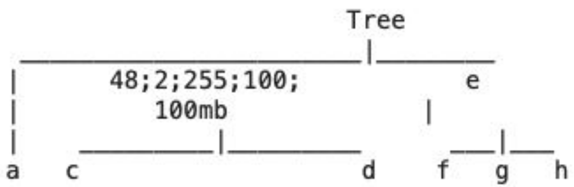
- couldn’t print color/bold text



```
                  Tree
  ___________________|__________
 |       b           e
 |    _______|________      ___|___
 a   c                d    f   g   h
```



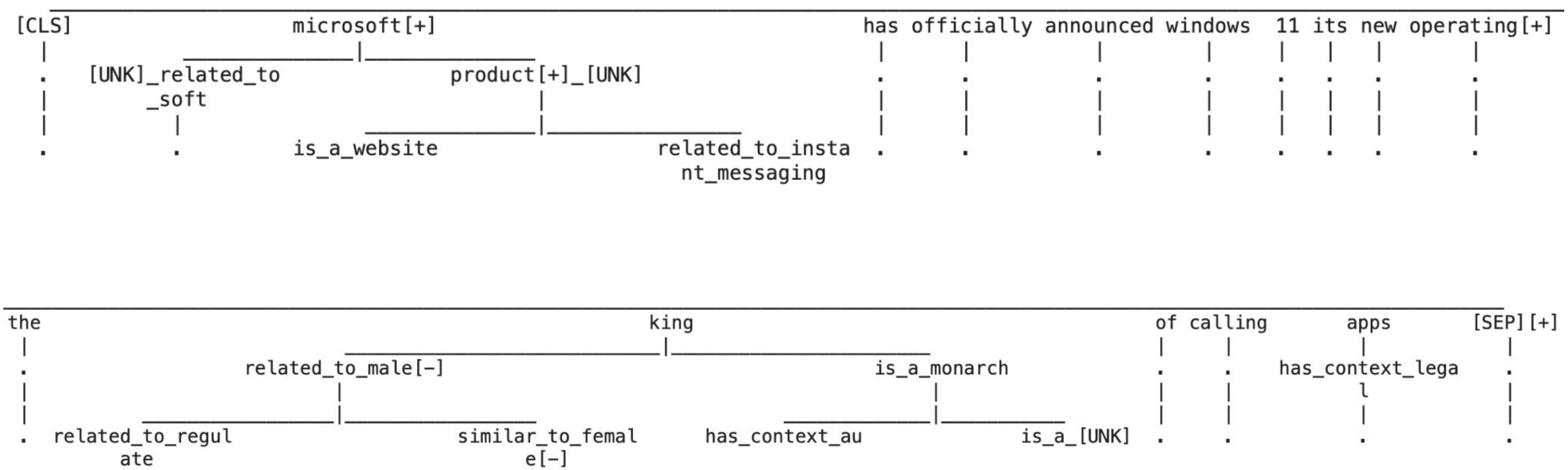
```
                  Tree
  ___________________|__________
 |       b           e
 |    _______|_______      ___|___
 a   c               d    f   g   h
```



```
                          Tree
  _________________________|_________
 |      48;2;255;100;               e
 |         100mb                    |
 |    __________|__________      ___|___
 a   c                     d    f   g   h
```

Results

nltk.tree (tech news)



```
   ______________________________________________________________________________________________________________________________
[CLS]                  microsoft[+]                                   has officially announced windows  11 its new operating[+]
  |           ______________|______________                            |       |          |        |     |   |   |       |
  .   [UNK]_related_to              product[+]_[UNK]                   .       .          .        .     .   .   .       .
  |        _soft                           |                           |       |          |        |     |   |   |       |
  |          |               ______________|________________           |       |          |        |     |   |   |       |
  .          .         is_a_website                  related_to_insta  .       .          .        .     .   .   .       .
                                                       nt_messaging
```

```
_____________________________________________________________________________________________________________________________
the                                                   king                                      of calling      apps      [SEP][+]
 |                           __________________________|_____________________                   |     |          |           |
 .                  related_to_male[-]                                  is_a_monarch            .     .   has_context_lega   .
 |                          |                                                |                  |     |          l           |
 |          ________________|________________                    ____________|___________       |     |          |           |
 .  related_to_regul                  similar_to_femal    has_context_au             is_a_[UNK] .     .          .           .
          ate                              e[-]
```

Results

Word Attribution Scores

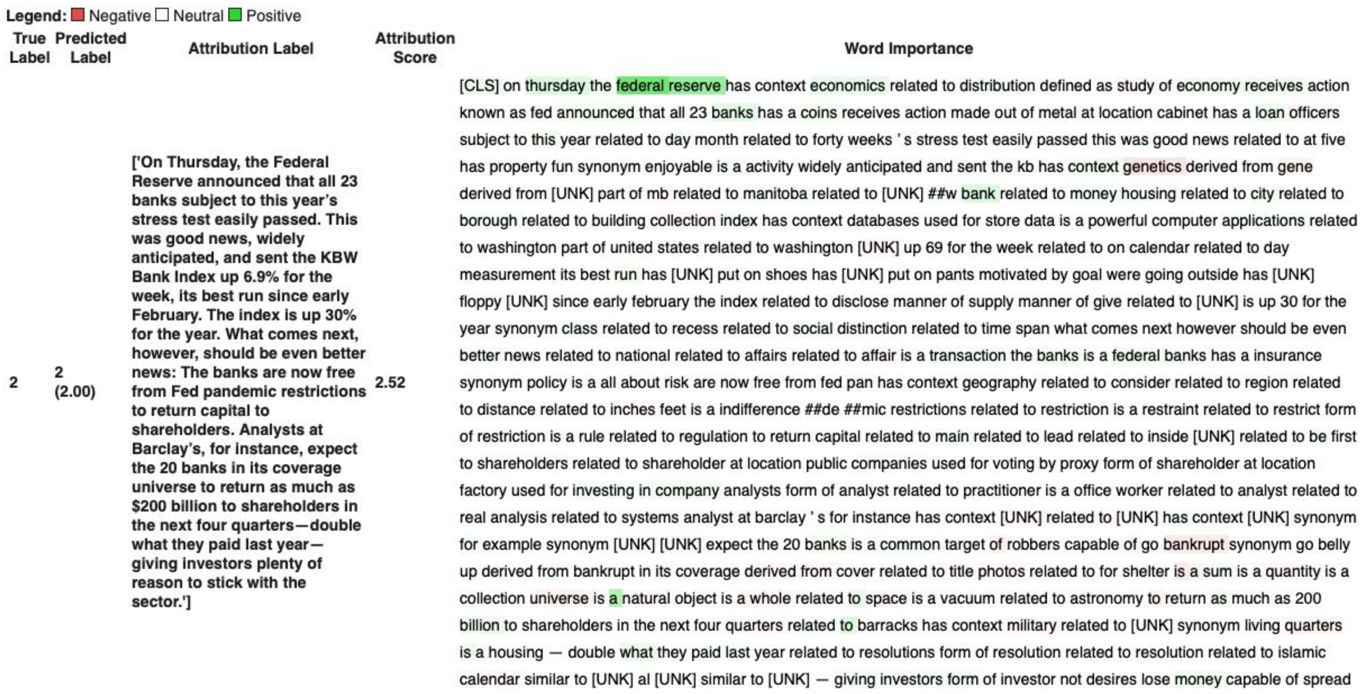
	original sentence	knowledge graph
positive	3.6458	4.8810
negative	-0.4748	-0.9292

Tokens

	original sentence	knowledge graph
positive	microsoft, announced, windows, new, operating, system, teams	to, a, company, enterprise, used, for, business, new
negative	king	Sea, women, child, mother, male

Results

Captum Visualization (business news)

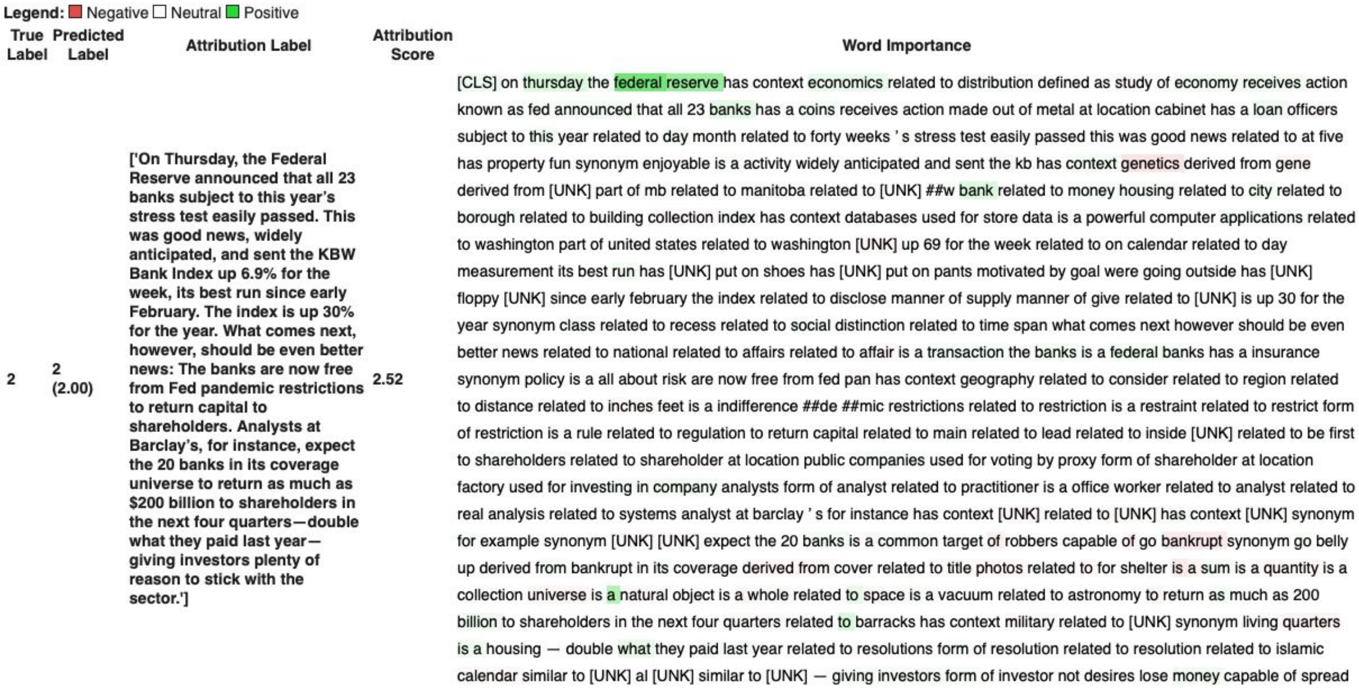


Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
2	2 (2.00)	['On Thursday, the Federal Reserve announced that all 23 banks subject to this year's stress test easily passed. This was good news, widely anticipated, and sent the KBW Bank Index up 6.9% for the week, its best run since early February. The index is up 30% for the year. What comes next, however, should be even better news: The banks are now free from Fed pandemic restrictions to return capital to shareholders. Analysts at Barclay's, for instance, expect the 20 banks in its coverage universe to return as much as $200 billion to shareholders in the next four quarters—double what they paid last year—giving investors plenty of reason to stick with the sector.']	2.52	[CLS] on thursday the federal reserve has context economics related to distribution defined as study of economy receives action known as fed announced that all 23 banks has a coins receives action made out of metal at location cabinet has a loan officers subject to this year related to day month related to forty weeks ' s stress test easily passed this was good news related to at five has property fun synonym enjoyable is a activity widely anticipated and sent the kb has context genetics derived from gene derived from [UNK] part of mb related to manitoba related to [UNK] ##w bank related to money housing related to city related to borough related to building collection index has context databases used for store data is a powerful computer applications related to washington part of united states related to washington [UNK] up 69 for the week related to on calendar related to day measurement its best run has [UNK] put on shoes has [UNK] put on pants motivated by goal were going outside has [UNK] floppy [UNK] since early february the index related to disclose manner of supply manner of give related to [UNK] is up 30 for the year synonym class related to recess related to social distinction related to time span what comes next however should be even better news related to national related to affairs related to affair is a transaction the banks is a federal banks has a insurance synonym policy is a all about risk are now free from fed pan has context geography related to consider related to region related to distance related to inches feet is a indifference ##de ##mic restrictions related to restriction is a restraint related to restrict form of restriction is a rule related to regulation to return capital related to main related to lead related to inside [UNK] related to be first to shareholders related to shareholder at location public companies used for voting by proxy form of shareholder at location factory used for investing in company analysts form of analyst related to practitioner is a office worker related to analyst related to real analysis related to systems analyst at barclay ' s for instance has context [UNK] related to [UNK] has context [UNK] synonym for example synonym [UNK] [UNK] expect the 20 banks is a common target of robbers capable of go bankrupt synonym go belly up derived from bankrupt in its coverage derived from cover related to title photos related to for shelter is a sum is a quantity is a collection universe is a natural object is a whole related to space is a vacuum related to astronomy to return as much as 200 billion to shareholders in the next four quarters related to barracks has context military related to [UNK] synonym living quarters is a housing — double what they paid last year related to resolutions form of resolution related to resolution related to islamic calendar similar to [UNK] al [UNK] similar to [UNK] — giving investors form of investor not desires lose money capable of spread

Results

Captum Visualization (business news)



Legend: 🟥 Negative □ Neutral 🟩 Positive

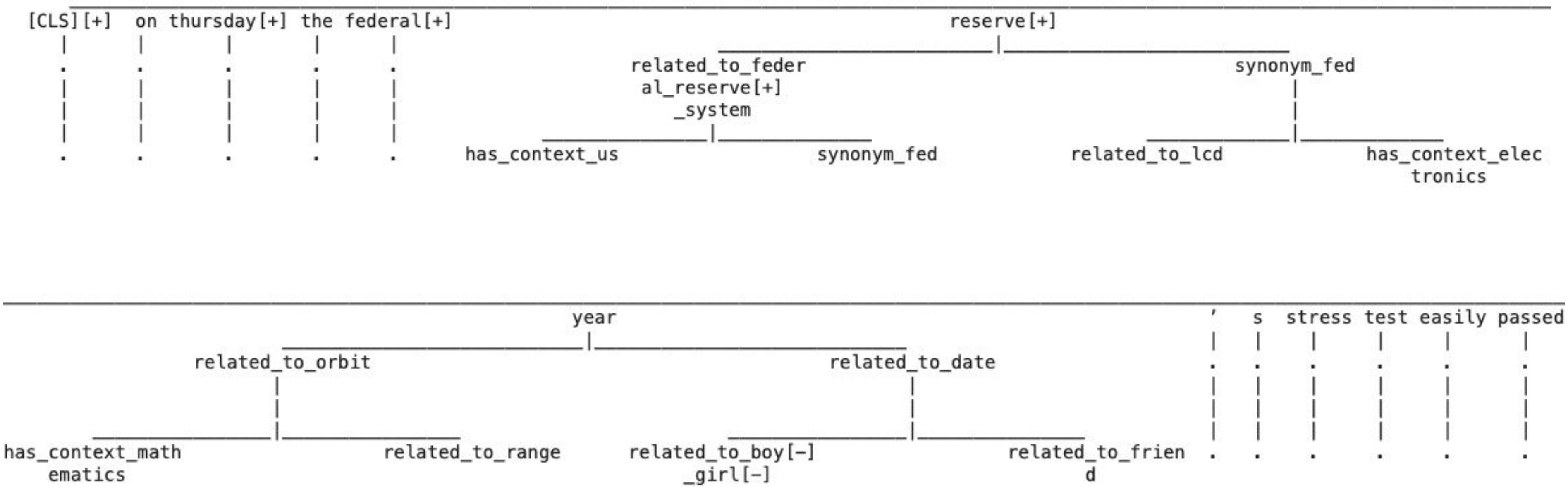
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
2	2 (2.00)	['On Thursday, the Federal Reserve announced that all 23 banks subject to this year's stress test easily passed. This was good news, widely anticipated, and sent the KBW Bank Index up 6.9% for the week, its best run since early February. The index is up 30% for the year. What comes next, however, should be even better news: The banks are now free from Fed pandemic restrictions to return capital to shareholders. Analysts at Barclay's, for instance, expect the 20 banks in its coverage universe to return as much as $200 billion to shareholders in the next four quarters—double what they paid last year—giving investors plenty of reason to stick with the sector.']	2.52	[CLS] on thursday the federal reserve has context economics related to distribution defined as study of economy receives action known as fed announced that all 23 banks has a coins receives action made out of metal at location cabinet has a loan officers subject to this year related to day month related to forty weeks ' s stress test easily passed this was good news related to at five has property fun synonym enjoyable is a activity widely anticipated and sent the kb has context genetics derived from gene derived from [UNK] part of mb related to manitoba related to [UNK] ##w bank related to money housing related to city related to borough related to building collection index has context databases used for store data is a powerful computer applications related to washington part of united states related to washington [UNK] up 69 for the week related to on calendar related to day measurement its best run has [UNK] put on shoes has [UNK] put on pants motivated by goal were going outside has [UNK] floppy [UNK] since early february the index related to disclose manner of supply manner of give related to [UNK] is up 30 for the year synonym class related to recess related to social distinction related to time span what comes next however should be even better news related to national related to affairs related to affair is a transaction the banks is a federal banks has a insurance synonym policy is a all about risk are now free from fed pan has context geography related to consider related to region related to distance related to inches feet is a indifference ##de ##mic restrictions related to restriction is a restraint related to restrict form of restriction is a rule related to regulation to return capital related to main related to lead related to inside [UNK] related to be first to shareholders related to shareholder at location public companies used for voting by proxy form of shareholder at location factory used for investing in company analysts form of analyst related to practitioner is a office worker related to analyst related to real analysis related to systems analyst at barclay ' s for instance has context [UNK] related to [UNK] has context [UNK] synonym for example synonym [UNK] [UNK] expect the 20 banks is a common target of robbers capable of go bankrupt synonym go belly up derived from bankrupt in its coverage derived from cover related to title photos related to for shelter is a sum is a quantity is a collection universe is a natural object is a whole related to space is a vacuum related to astronomy to return as much as 200 billion to shareholders in the next four quarters related to barracks has context military related to [UNK] synonym living quarters is a housing — double what they paid last year related to resolutions form of resolution related to resolution related to islamic calendar similar to [UNK] al [UNK] similar to [UNK] — giving investors form of investor not desires lose money capable of spread

thursday, federal, reserve, economics, economy, bank, coin, loan, city, billion

genetics, gene, of, bankrupt, universe, quarters

Results

nltk.tree (business news)



```
       ________________________________________________________________________________________________________________________________________
   [CLS][+]  on thursday[+] the federal[+]                                             reserve[+]
       |      |       |       |      |                            __________________________|__________________________
       .      .       .       .      .                    related_to_feder                                       synonym_fed
       |      |       |       |      |                     al_reserve[+]                                               |
       |      |       |       |      |                        _system                                                  |
       |      |       |       |      |            ________________|______________                        ______________|_____________
       .      .       .       .      .     has_context_us                  synonym_fed            related_to_lcd             has_context_elec
                                                                                                                                 tronics
```

```
 ______________________________________________________________________________________________________________________________________________
                                                     year                                                      '   s  stress test easily passed
                           ___________________________|_____________________________                           |   |    |      |     |      |
                   related_to_orbit                                         related_to_date                    .   .    .      .     .      .
                          |                                                         |                          |   |    |      |     |      |
                          |                                                         |                          |   |    |      |     |      |
         _________________|________________                        _________________|_______________           |   |    |      |     |      |
 has_context_math                   related_to_range      related_to_boy[-]                  related_to_frien  .   .    .      .     .      .
     ematics                                                   _girl[-]                             d
```

Results

Word Attribution Scores

	original sentence	knowledge graph
positive	2.8145	3.2272
negative	-0.5607	-2.9578

Tokens

	original sentence	knowledge graph
positive	Thursday, federal, reserve, banks, billion	economics, economy, coins, loan, city,
negative	universe, quarters	genetics, gene, of, bankrupt

Our Multi-hop K-bert Approach

Advantages

- Better explainability
- Broad knowledge
- Easy interpretation

Limitations

- Limited number of tokens (Bert model)
- Irrelevant neighbours

Future Works

- Use cosine similarity of embeddings to add relevant neighbours
  - Sentence - word similarity
  - Word - word similarity

- Downside
  - Can’t use multi-word tokens
  - Slow
  - Inefficient
    - Better to use an embedding-using model

Thank you