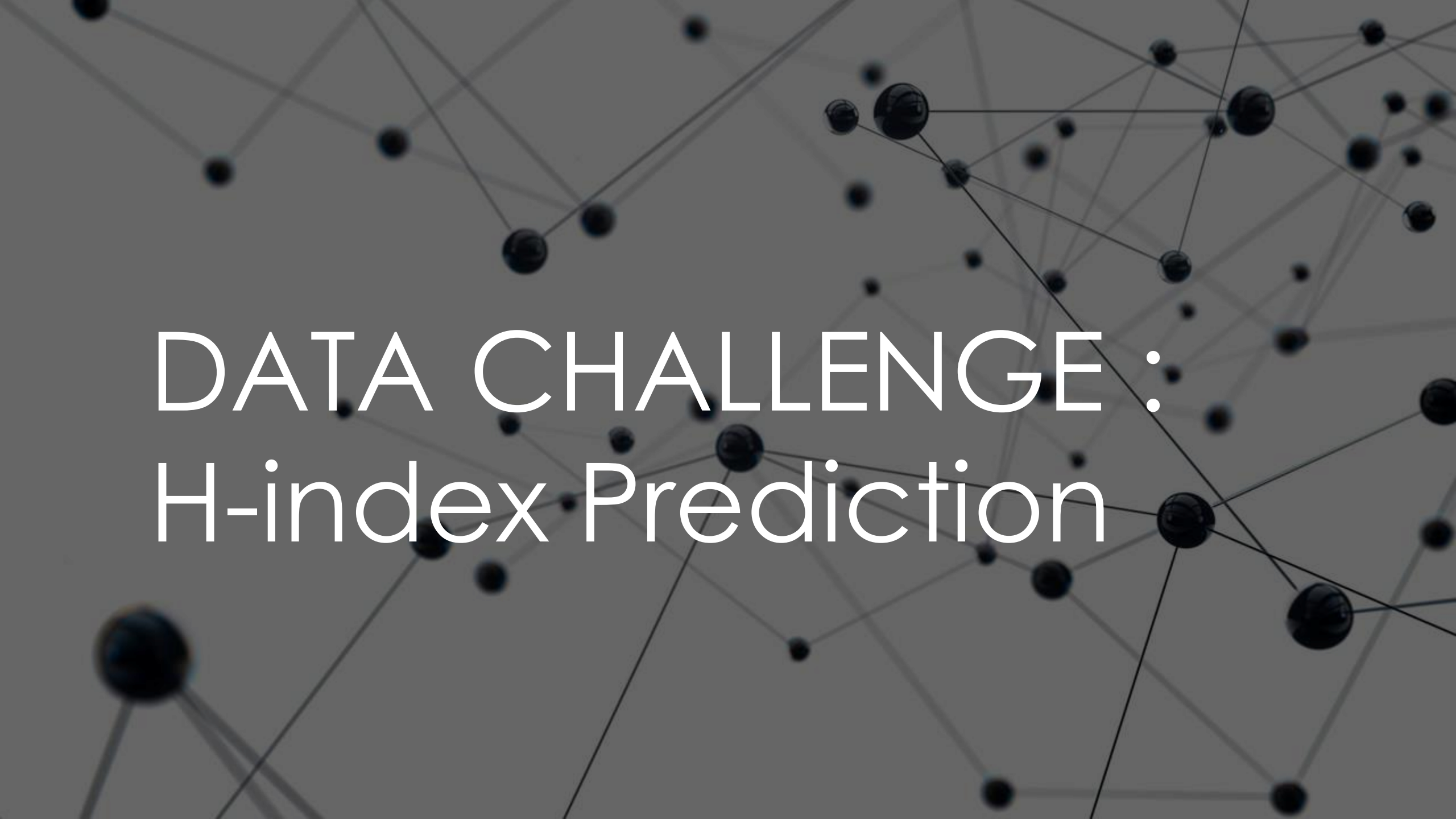


DATA CHALLENGE : H-index Prediction

Plan

- Introduction
- Features Extraction
- Model Tuning and Results
- Conclusion

Introduction

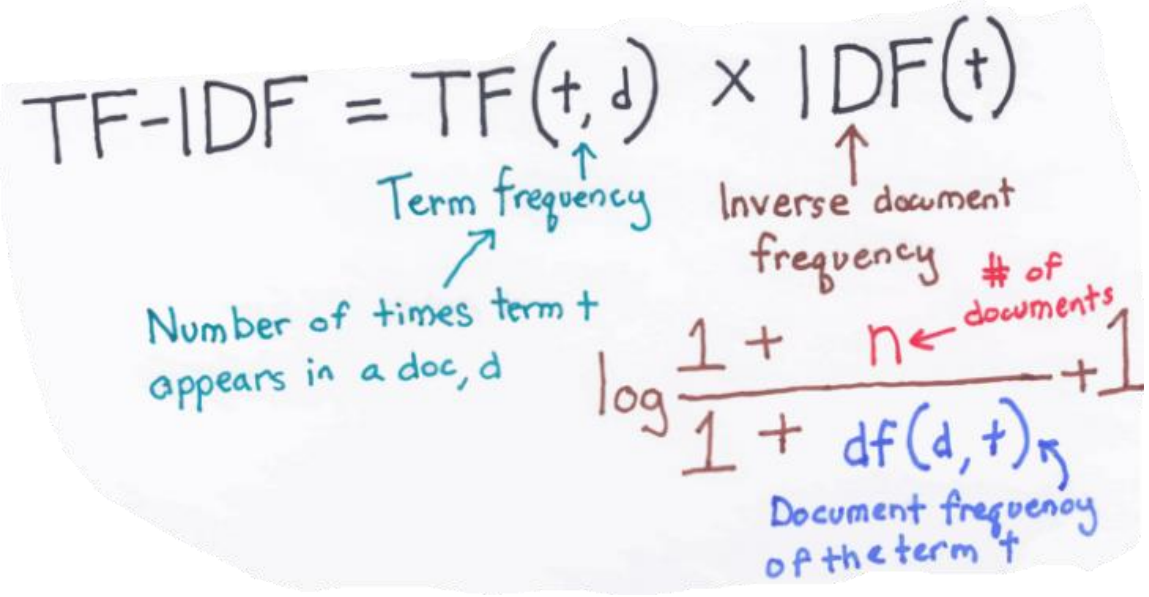
- The aim of our work was to predict the h-index of an researcher
- The h-index an researcher measures his/her productivity and the citation impact of his/her Publications.
- This h-index can be defined as the maximum value of h such that the given author has published h papers that have each been cited at least h times.
- The data provided for this project were a graph of collaboration between researchers and the abstracts of the top cited papers of each researchers
- For our project we have followed the general pipeline of an ML project which consist of firstly, extract features from our Raws data which can be directly pass to an ML model, test several model base on these features and fine-tuning ours models in order to improve the performance of this one.

Features Extraction

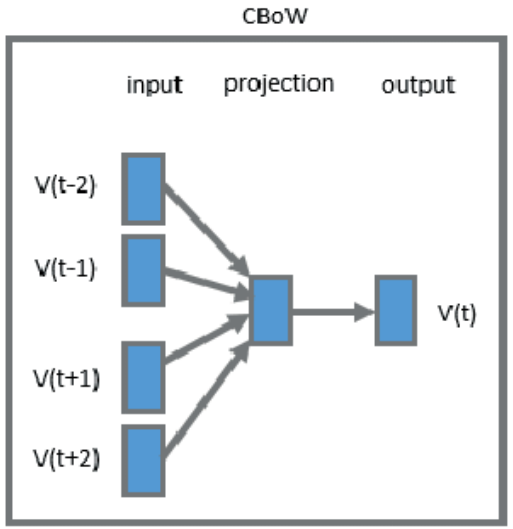
Abstract features

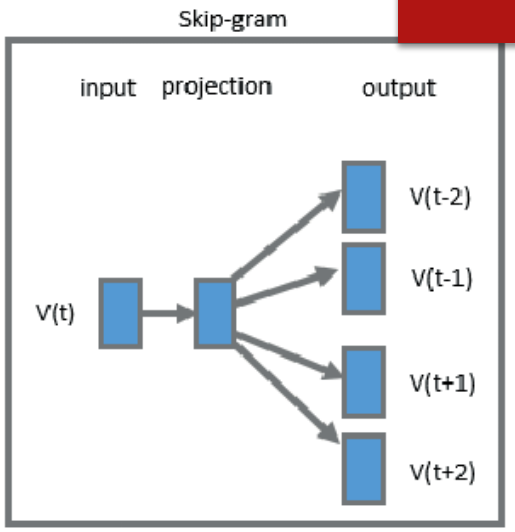
- Strategy : TFIDF weighted Word2Vec

- TF-IDF Term Frequency-Inverse Document Frequency : is a measure originality of a word by comparing the number of times a word appears in a document with the number of docs the word appears in.
- Implementation: tfidfVectorizer() from scikitlearn

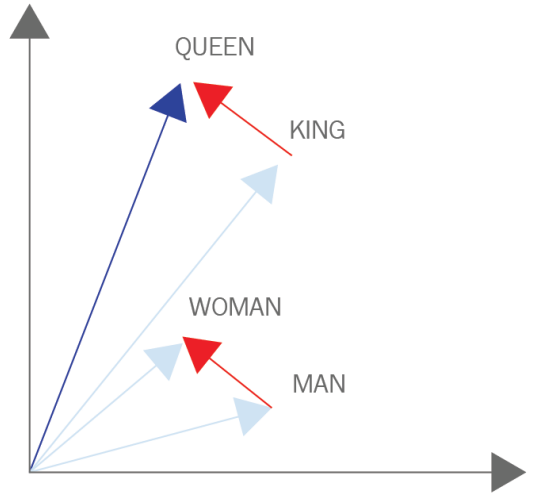


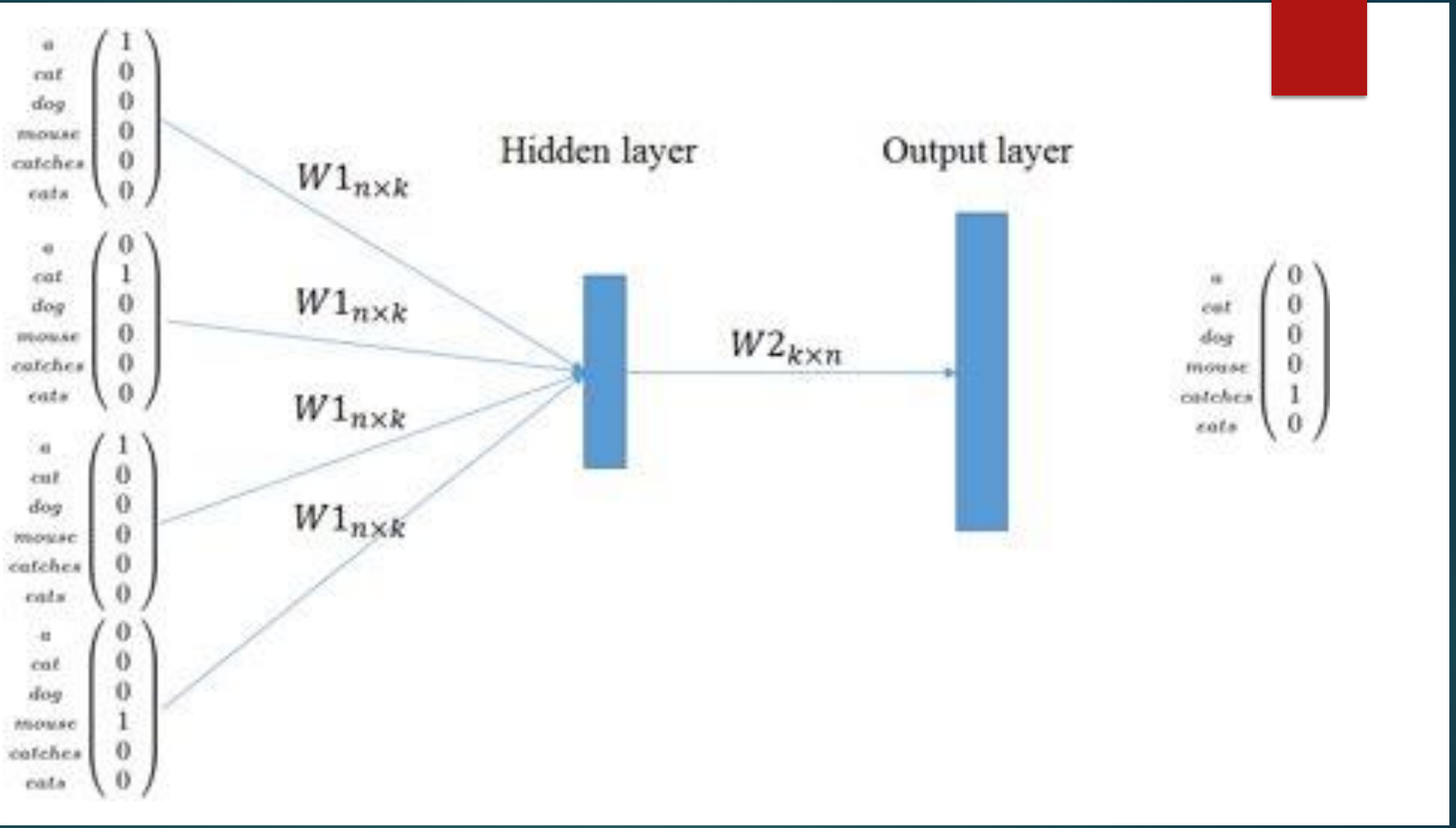
- CBOW (Continuous Bag-Of-Words) or Skip-gram
- Word2Vec: Is a technique for NLP published in 2013.





So king + man - woman = queen!




a
cat
dog
mouse
catches
eats
1 0 0 0 0 0
0 1 0 0 0 0
1 0 0 0 0 0
0 0 0 1 0 0
$W1_{n\times k}$
$W1_{n\times k}$
$W1_{n\times k}$
$W1_{n\times k}$
Hidden layer
$W2_{k\times n}$
Output layer
0 0 0 0 1 0

Preprocessing of abstracts

- **abstract_sentenses for TFIDF :** "in this paper we describe a new bitmap indexing technique to cluster xml documents xml is a new standard for exchanging and representing information"

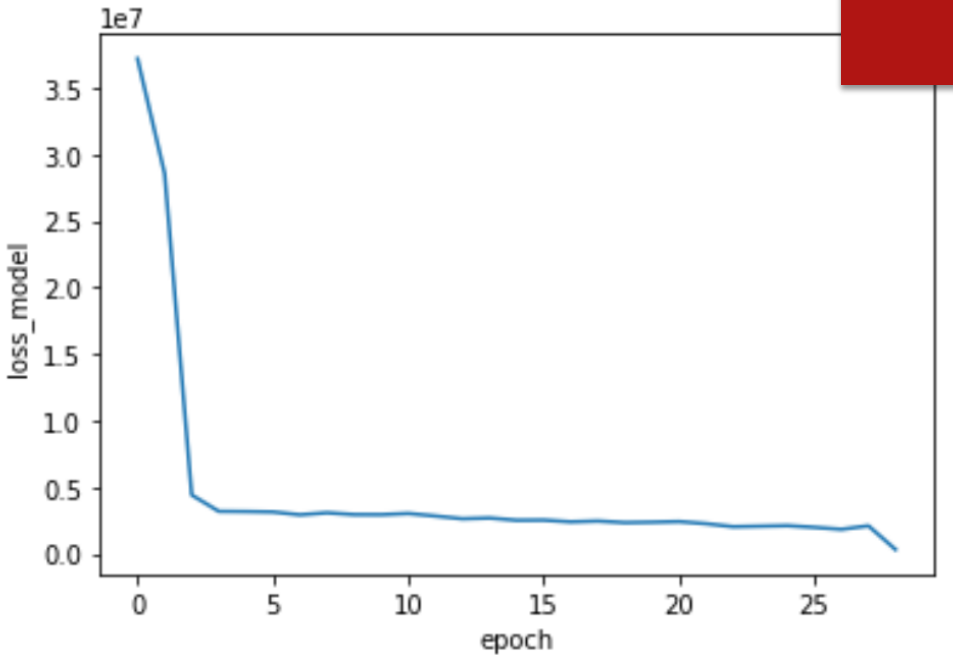
- **abstract_word_table for word2vec :**['in', 'this', 'paper', 'we', 'describe', 'a', 'new', 'bitmap', 'indexing', 'technique', 'to', 'cluster', 'xml', 'documents', 'xml', 'is', 'a', 'new', 'standard', 'for', 'exchanging', 'and', 'representing', 'information']

Number of abstracts : 624168

Vocabulary size : 427456

Word2vec Training : 7 hours of training

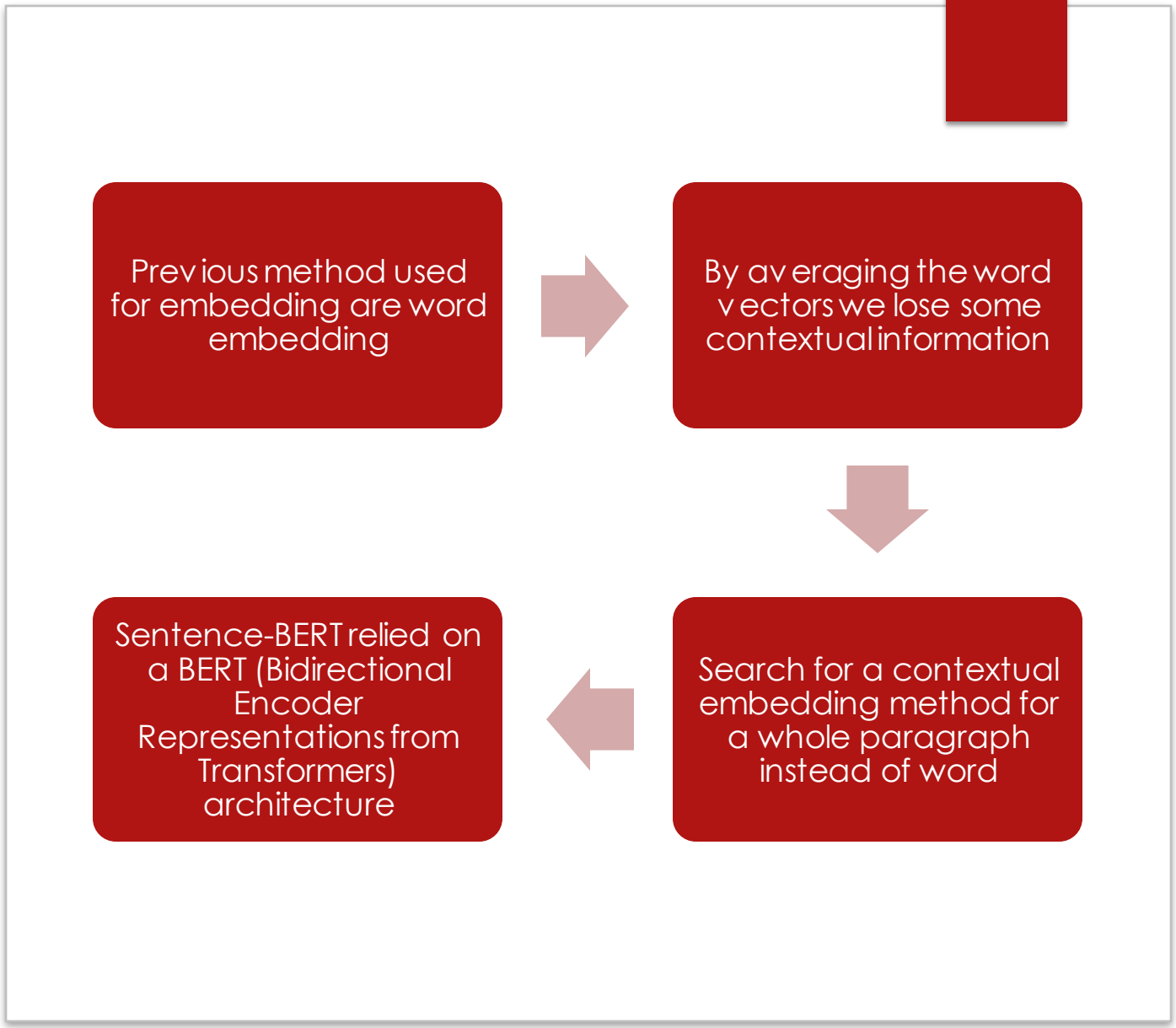
- 'Taikang' is Chinese company. Our model found well that it is related to company.
- The usual word2Vec doesn't contain it.




```
1 reload_model.wv.most_similar(positive='company', to
```

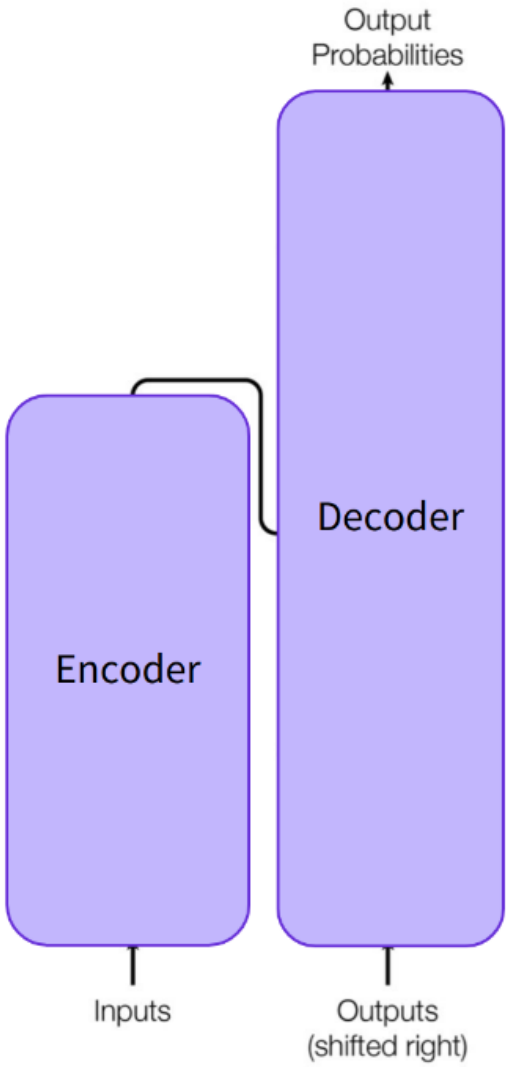
```
[('companies', 0.6483214497566223),
 ('customers', 0.6270943880081177),
 ('customer', 0.601658046245575),
 ('taikang', 0.5920459628105164),
 ('company's', 0.5890681743621826),
 ('utopics', 0.5839164853096008),
 ('organizations', 0.5789316892623901),
 ('employees', 0.5730001926422119),
 ('financial', 0.5720323324203491),
 ('bangchak', 0.5690685510635376)]
```

Sentence-BERT



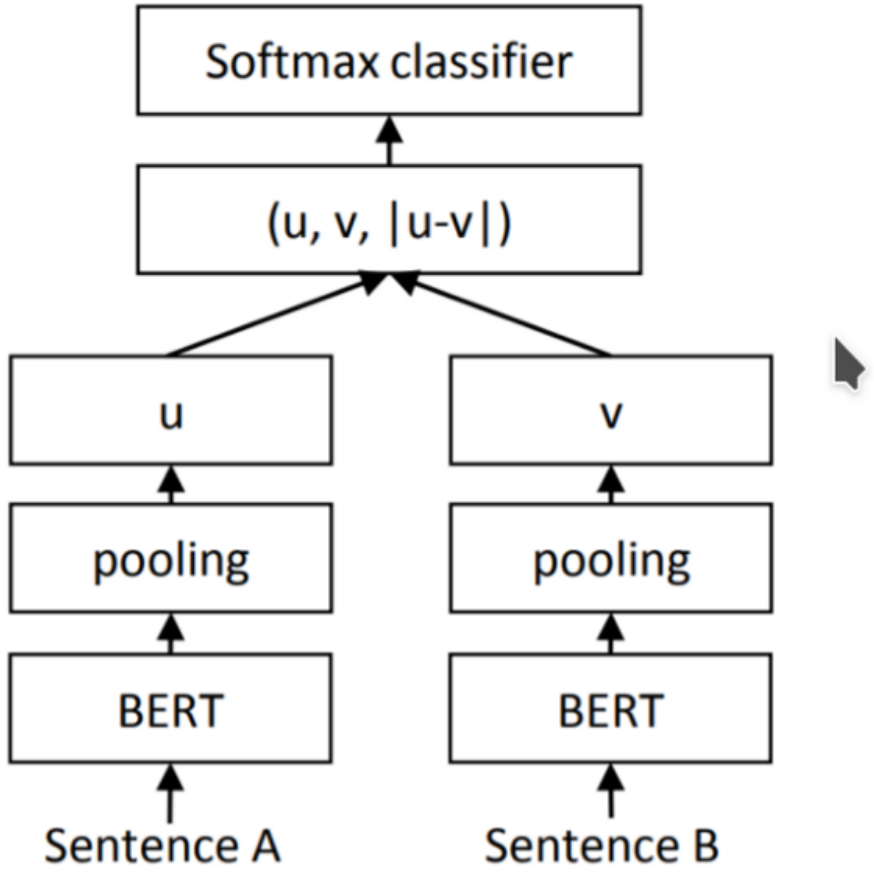
BERT and Transformers

- ▶Encoder : The encoder receives an input and builds a representation of it (features). This means that the model is optimized to acquire understanding from the input.
- ▶Decoder : The decoder uses the encoder's representation (features) along with other inputs to generate a target sequence. This means that the model is optimized for generating outputs.
- ▶these parts can be used independently, depending on the task
- ▶BERT is a Encoder-only models



How does Sentence-BERT work?

- The idea is to fine-tune BERT sentence embeddings on a dataset which rewards models that generates sentence embeddings that have the following property:
- When the cosine similarity of the pair of sentence embeddings is computed, we want it to represent accurately the semantic similarity of the two sentences.



Graph Features

Graph metrics:

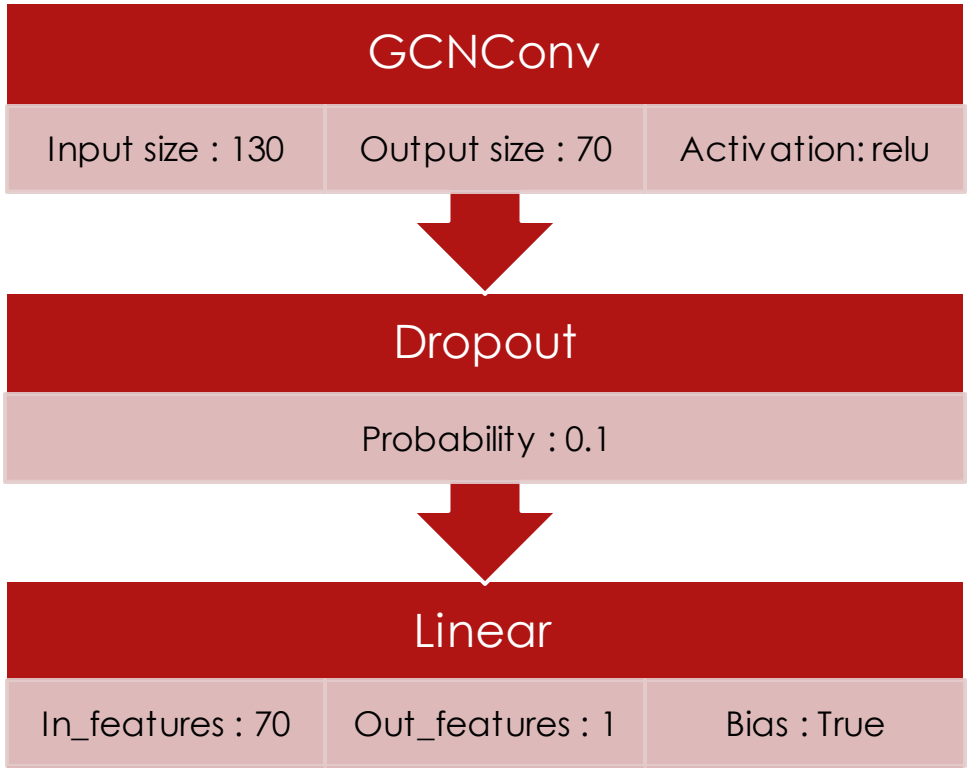
- Degree: The sum of the weights of edges adjacent to a vertex.
- Degree centrality: It is the normalized degree of a vertex
- Neighbor's average degree: The average degree of the neighborhood of a vertex
- PageRank: Pagerank is an algorithm that computes a ranking of the vertices in a graph based on the structure of the incoming egdes
- Core number: A subgraph of a graph G is defined to be a k-core of G if it is a maximal subgraph of G in which all vertices have degree at least k.
- Onion layers: The onion decomposition is a variant of the k-core decomposition
- Diversity coefficient: The diversity coefficient is a centrality measure based on the Shannon entropy
- Community-based centrality: This centrality measure calculates the importance
- of a vertex by considering its edges towards the different communities .

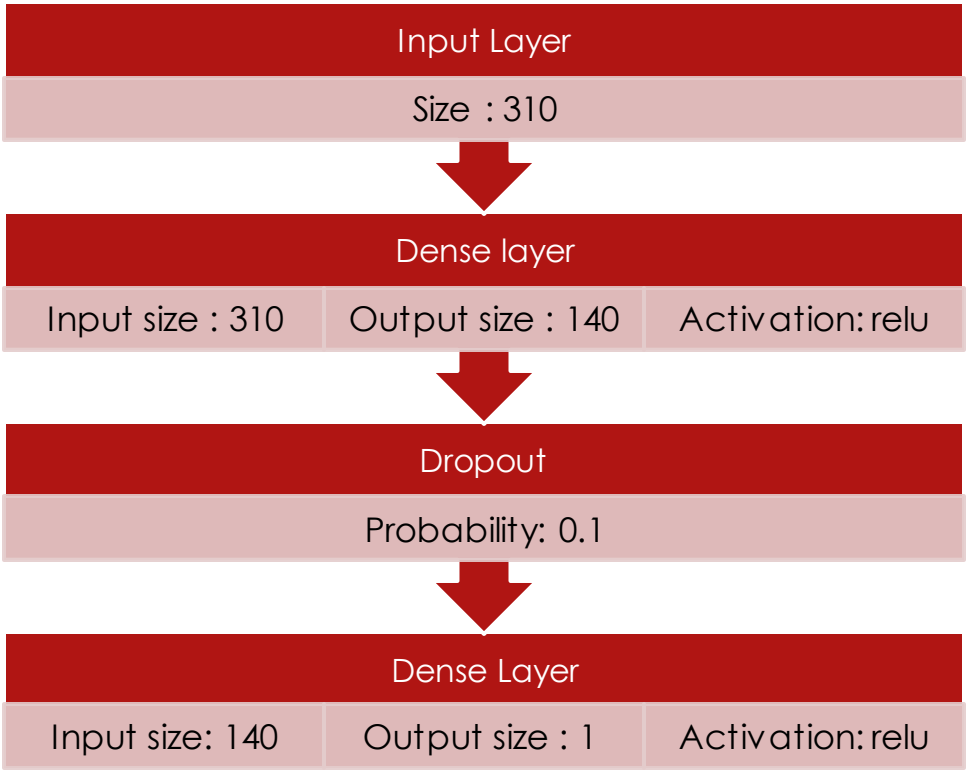
Node2Vec(G, dimensions=100, walk_length=30, num_walks=200, workers=4)

MODEL TUNING & RESULT

Model Tuning

GCN and MLP





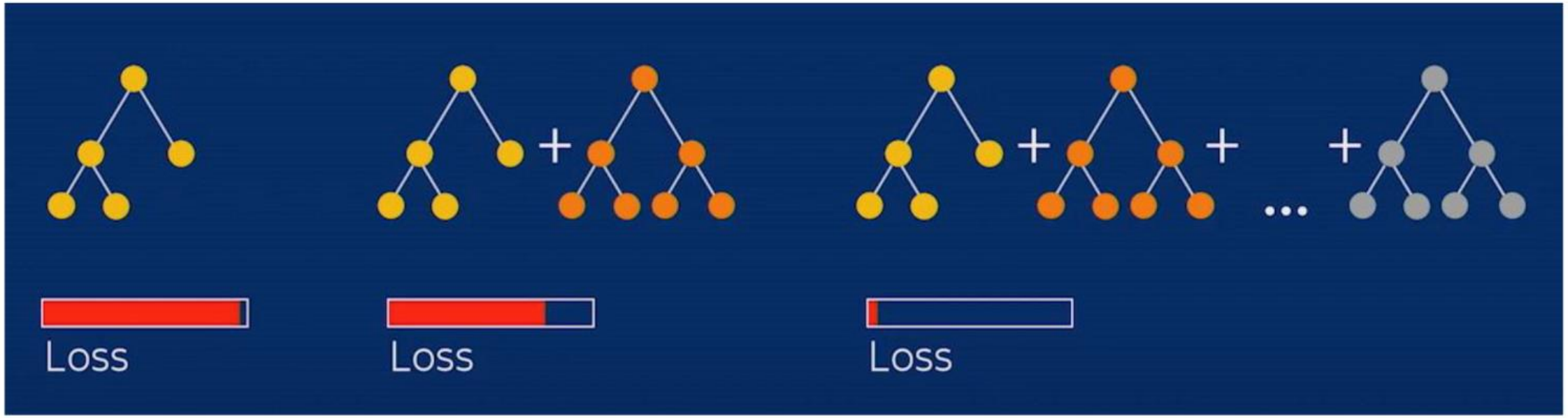
XGBoost

Gradient boosting machine learning algorithm

It is an ensemble of multiple weak learners generaly decision tree algorithms

In gradient boosting, we built trees step by step and add them to the composition

At each step we look for a basic algorithm that corrects the composition error in the previous step


Loss
Loss
Loss

Results

method	MSE with Text features	MSE with Graph features	MSE with all features
XGBoostRgressor	75,45	77,12	53,40
MLPRegressor	71,88	81,53	52,72
GCNRegressor	78,24	67,34	56,43

Conclusion