Data Mining & Machine Learning Project

Resume Classification

Daniele Laporta

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

CROSSLAB
Innovation for industry 4.0

Project steps

1. **Design:** Dataset description, preprocessing, EDA and feature transformation, sentiment score, word embeddings.
2. **Model Implementation:** Resume binary classification (STEM/ not STEM) with stratified k-fold cross validation.
3. **Classification Results**
4. **Conclusions**
5. **Possible Improvements**
6. **Additional study**

Developed with python™ **on** colab .

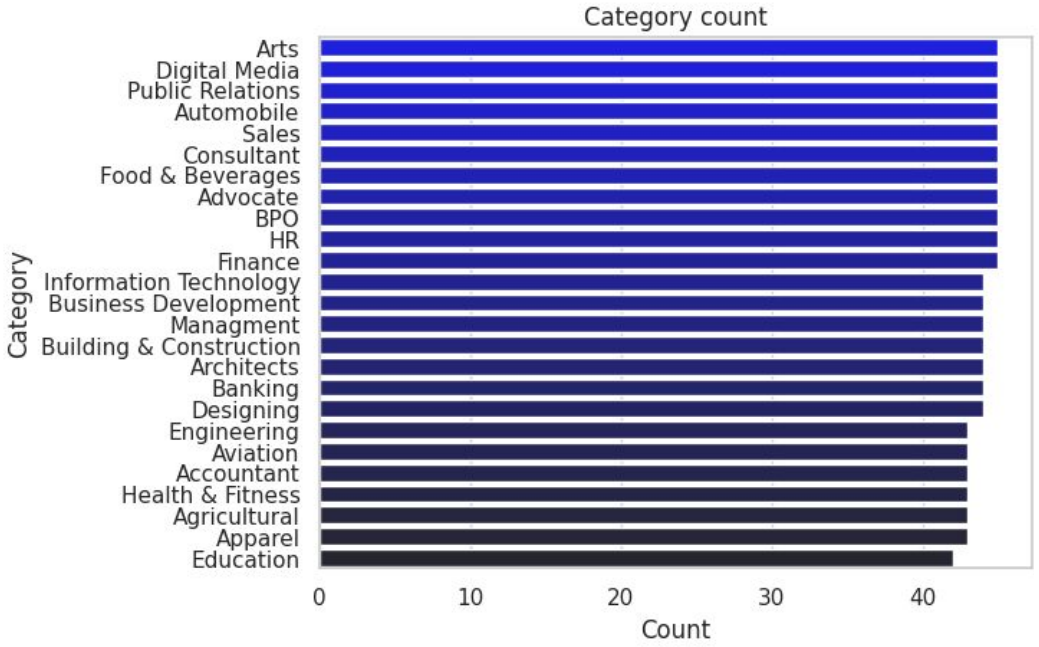
Dataset Description

**Resumes dataset from [Kaggle](): **
real people linkedin resumes publicly available. Each resume has many categorical features and a ground truth, ‘category’, which is the job sector.

category	linkedin	profile_picture	description	Experience	Name	position	location	skills	clean_skills
HR	https://in.linke...	https://media-ex...	An experienced H...	Senior Vice Pres...	Sameer Wadhawan	Senior Vice Pres...	Gurgaon, Haryana...	['\nPerformance ...	['Performance Ma...
HR	https://in.linke...	https://media-ex...	Head Talent Acqu...	Head of Talent A...	Adarsh Krishna	Head Talent Acqu...	Pune, Maharashtr...	['\nTalent Acqui...	['Talent Acquisi...
HR	https://in.linke...	data:image/gif;b...	A Talent Acquisi...	Company NameIBM ...	Shrivas Mohit	HR@IBM	Bengaluru, Karna...	['\nHuman Resour...	['Human Resource...
HR	https://in.linke...	data:image/gif;b...	NaN	HR/Admin/Personn...	HR Hopes	HR	Pune Area, India	[]	[']
HR	https://in.linke...	https://media-ex...	Over 18 Years of...	Company NameEXLT...	Rakesh Kumar	Vice President -...	Central Delhi, D...	['\nTeam Managem...	['Team Managemen...
...	...	...	...	...	...	...	...	...	...
Aviation	https://in.linke...	data:image/gif;b...	NaN	airline and avia...	britishairhostes...	airline and avia...	Chandigarh, Chan...	[]	[']
Aviation	https://in.linke...	data:image/gif;b...	NaN	Production Manag...	Ramaiah Manjunath	Production Manag...	Krishnagiri, Tam...	[]	[']
Aviation	https://in.linke...	https://media-ex...	An MBA Graduate ...	Human Resources ...	Shubham Pradhan	HR Executive at ...	Mumbai, Maharash...	['\nManagement\n...	['Management', '...
Aviation	https://in.linke...	data:image/gif;b...	NaN	Institute of Log...	Alfiya Shaikh	Student at Insti...	Mumbai, Maharash...	['\nTally ERP\n'...	['Tally ERP', 'W...
Aviation	https://in.linke...	https://media-ex...	Currently I'm pu...	Executive Revenu...	ABHISHEK TIWARI	Pursuing MBA in ...	Thane, Maharasht...	['\nSix Sigma\n'...	['Six Sigma', 'M...

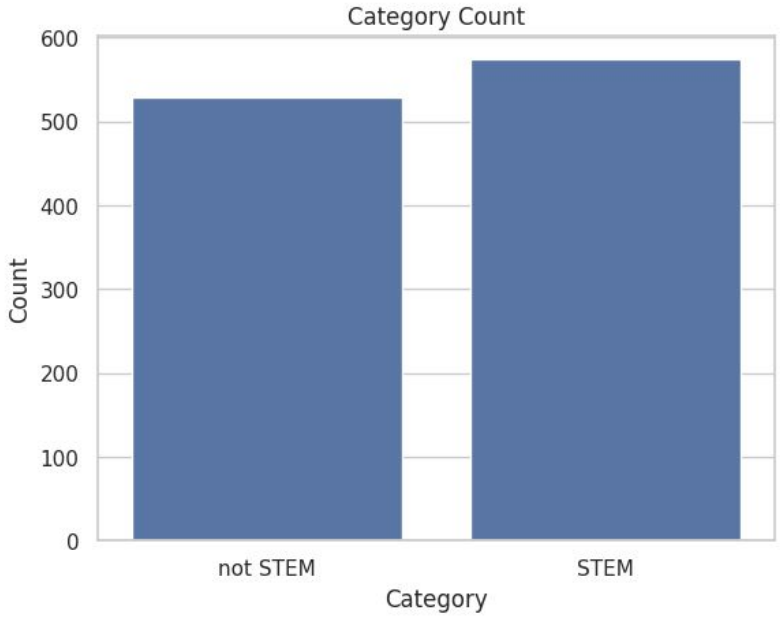
(Initial dataset: 1251 rowsx 11 columns → Final dataset: 1103 rows x 9 columns)

EDA



25 classes,
40 samples per class.
Not enough data to perform multi classification!

For simplicity, I reduced the ground truth to 2 classes:
STEM/ not STEM



DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

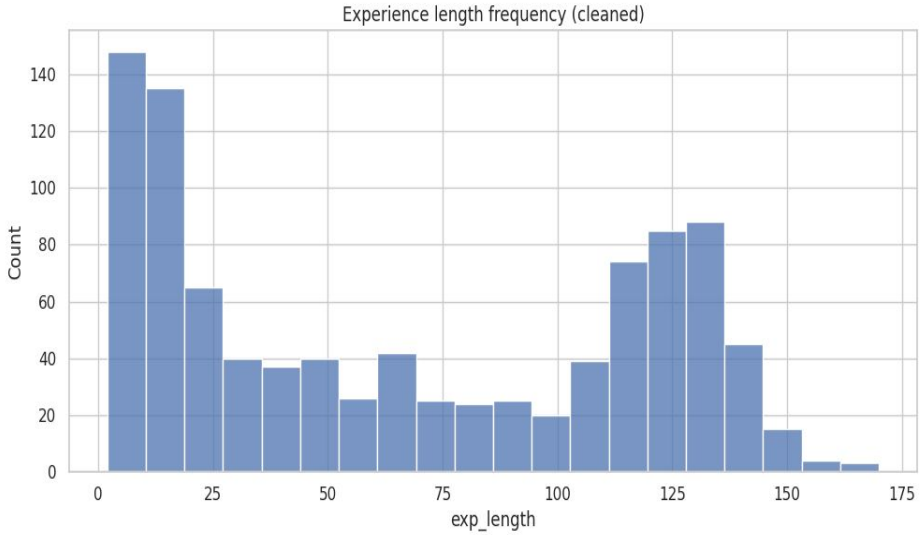
UNIVERSITÀ DI PISA

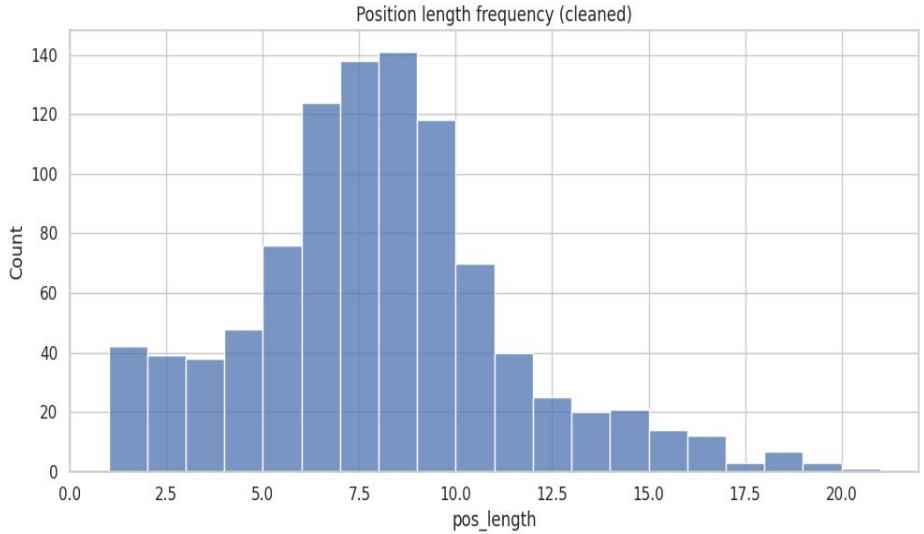
CROSSLAB Innovation for industry 4.0

EDA

‘skills’ Word Cloud

‘experience’ and ‘position’ length frequency



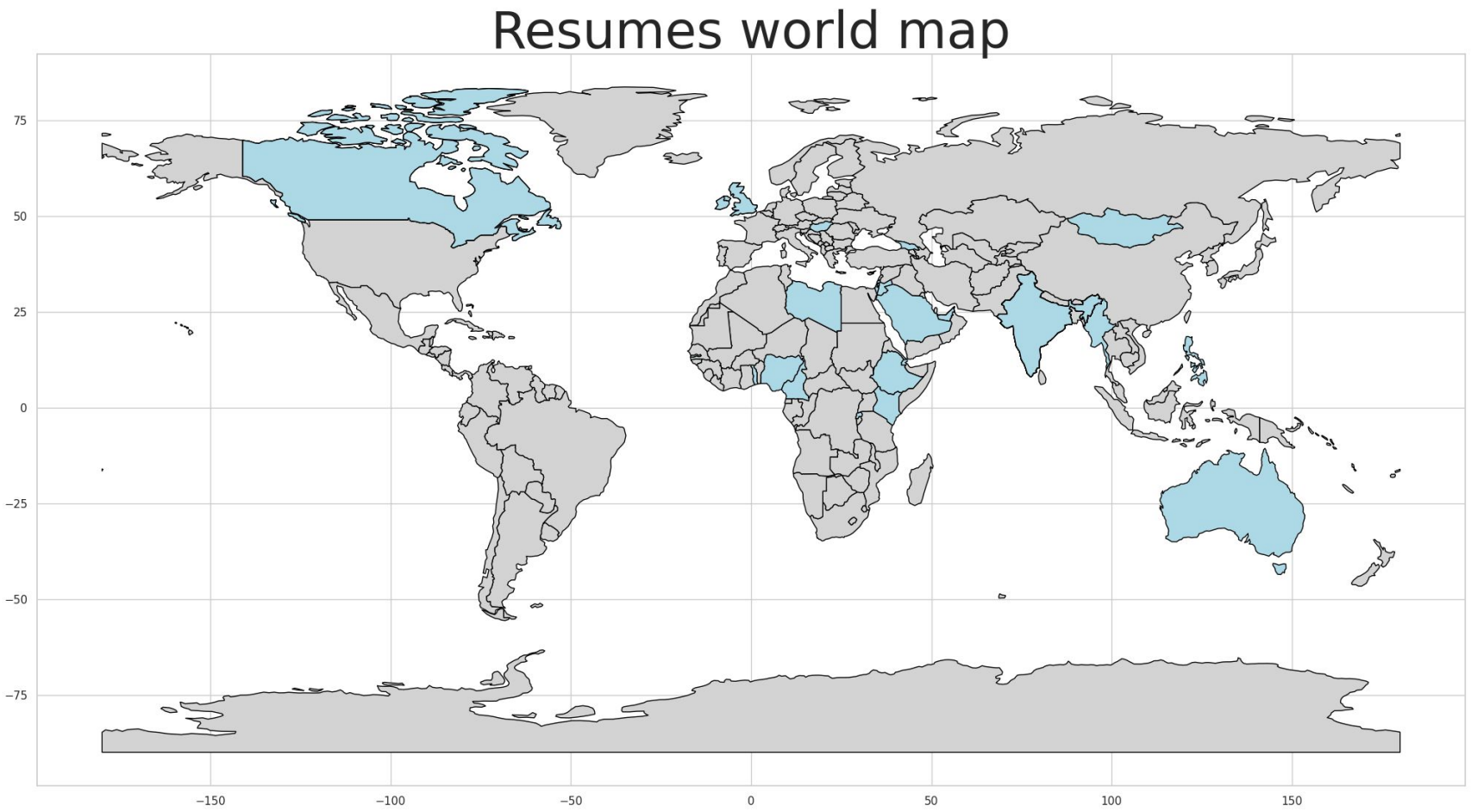


DIPARTIMENTO DI INGEGNERIA DELL’INFORMAZIONE

UNIVERSITÀ DI PISA

CROSSLAB Innovation for industry 4.0

EDA



* not all countries are recognised in the map.

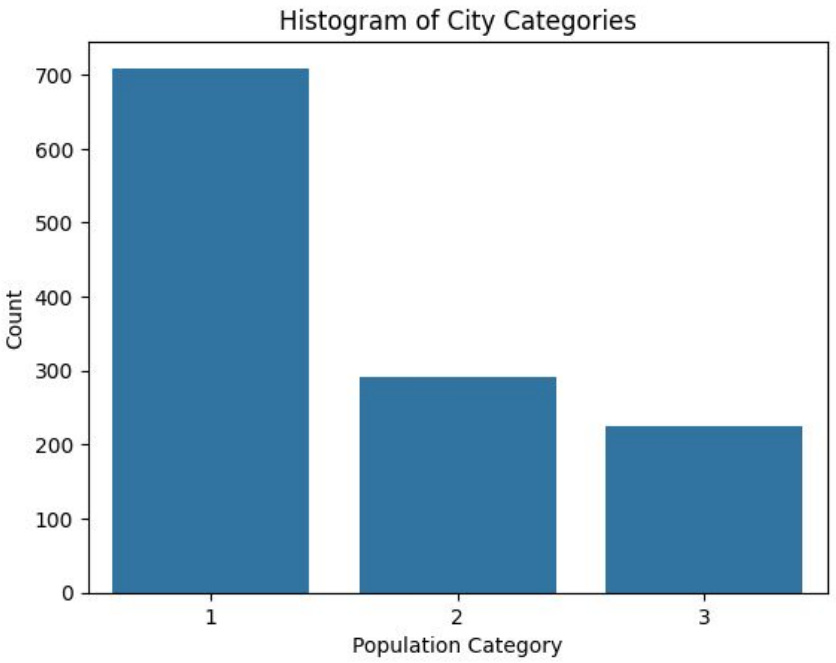
Feature Transformation

The following are performed on all the dataset.

- Transform ‘location’ column

I used a dataset (Geonames) of around 147k cities with their population size. The aim is to substitute the cities in my Kaggle dataset with a discretization of the population in terms of:

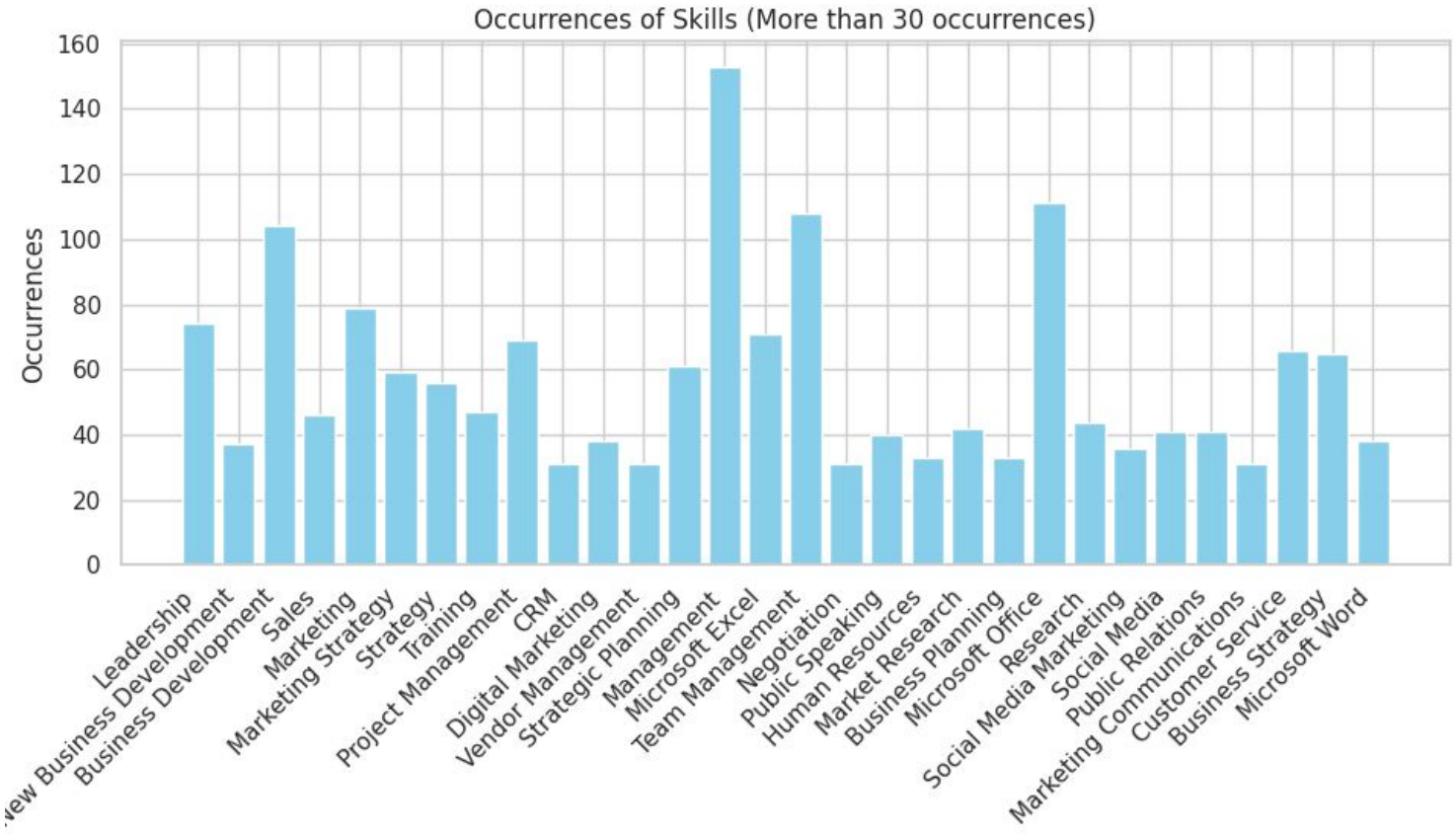
- population <1M: 'small' (1)
- 1M ≤population<5M: 'medium' (2),
- population ≥5M: 'big' (3).



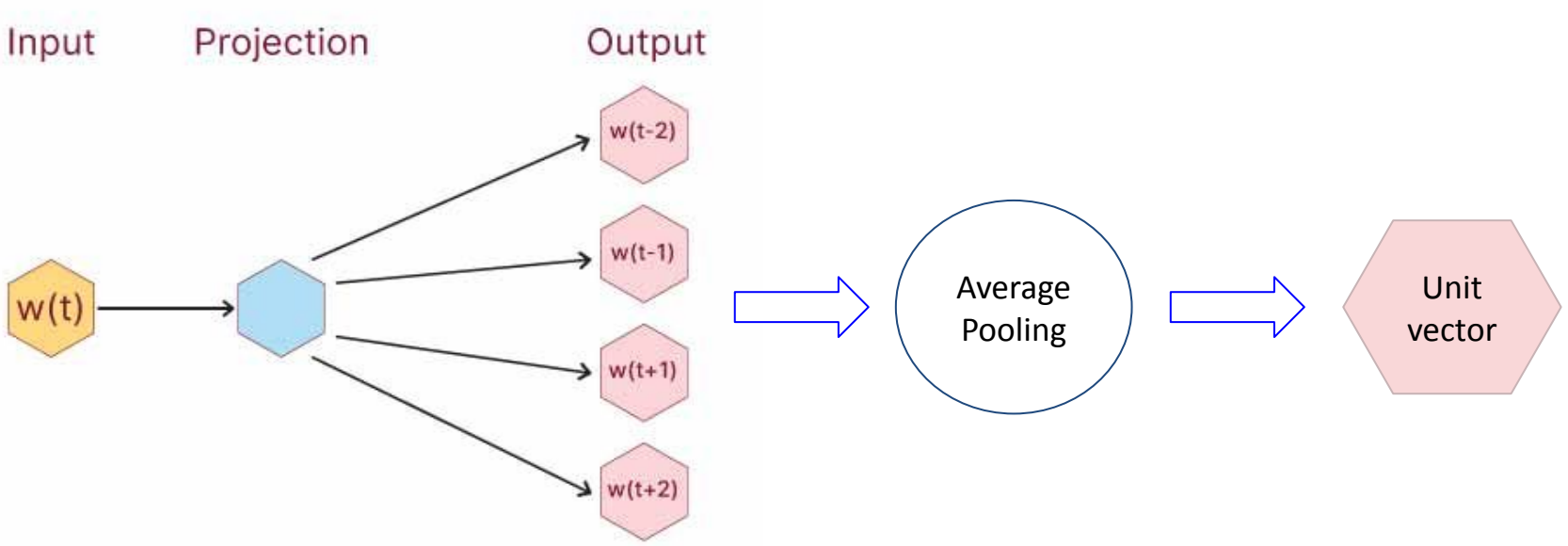
- Sentiment score for ‘experience’, ‘position’ and ‘skills’ columns using 2 pre-trained Hugging Face models, specific for job skills and descriptions.

Feature Transformation

- Transform (weight) 'skills' column: If a skill appears more times, it becomes less important for the model. A skill that appears very few times can be considered rare.



Word Embeddings



- Word2Vec, Doc2Vec and BERT for the ‘skills’ feature.
- Word2Vec, Doc2Vec for the ‘experience’ feature.
- Similar words ⟶ similar embeddings.

Classification Results

- Stratified K-Fold Cross Validation (K=5). 2 classes: STEM (1) / not STEM (0)
- Dataset basic configuration: 9 features: ‘exp_length', 'pos_length', 'population_category', ‘skills_scores_1’, ‘skills_scores_2’, 'position_scores_1', 'position_scores_2', 'experience_scores_1', 'experience_scores_2'.
- Dataset +3we configuration: same 9 features + 3 word embeddings on ‘skills’ feature.
- Dataset +5we configuration: +3we + 2 word embeddings on ‘experience’ feature.

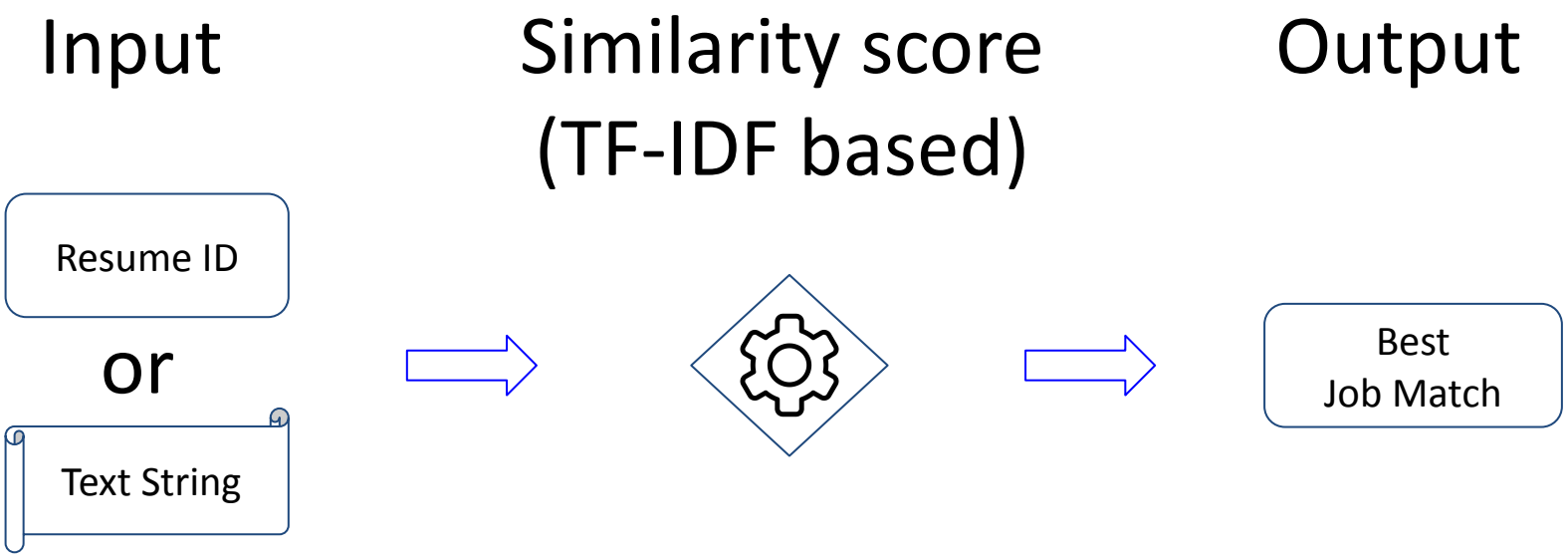
Dataset Configuration	Averaged metrics	LOG REG	RF	SVC	MULTIN. NB	XGBOOST
basic	Accuracy Precision Recall F1 Score	0.60 0.62 0.63 0.62	0.60 0.62 0.59 0.60	0.61 0.62 0.68 0.64	0.56 0.55 0.85 0.67	0.59 0.61 0.57 0.59
+3we		0.59 0.60 0.68 0.63	0.59 0.61 0.59 0.60	0.59 0.59 0.71 0.64	0.57 0.56 0.82 0.66	0.58 0.59 0.65 0.61
+5we		0.66 0.68 0.68 0.67	0.67 0.72 0.61 0.63	0.66 0.68 0.71 0.68	0.58 0.57 0.80 0.66	0.67 0.70 0.66 0.67

Possible Improvements

- Data augmentation or generally more data at disposal (E.g. salary, cost and quality of living).
- Focus on a niche job market (additional study).
- Study historical and evolving trends of required skills per job.
- Develop an application for resume-job matching.

Additional Study: Job Recommender

Source: Job postings dataset from Kaggle (data science and tech job offers)



Thank you for the attention!