BILSTM RECURRENT NEURAL NETWORKS IN HETEROGENEOUS ENSEMBLE MODELS FOR NAMED ENTITY RECOGNITION PROBLEM IN LONG POLISH UNSTRUCTURED DOCUMENTS

Aleksandra Świetlicka, Adam Iwaniak,
Mateusz Piwowarczyk

WROCŁAWSKI INSTYTUT ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI

WROCŁAWSKI INSTYTUT ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI

The Wroclaw Institute of Spatial Information and Artificial Intelligence

www.wizipisi.pl

WROCŁAWSKI INSTYTUT ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI

AGENDA

WROCŁAWSKI INSTYTUT ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI

AGENDA

1. Introduction
2. Data preparation
3. Methods
   - Artificial Neural Networks (ANNs)
   - Ensemble model with ANN, Random Forest and XGBoost
4. Results
5. Summary

INTRODUCTION

WROCŁAWSKI INSTYTUT ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI

PROBLEM STATEMENT

- Information extraction from long Polish documents with complex layouts
- Data to extract: name of the organization, address of the organization (separately city, street and street number), names of people managing the organization with their positions, and dates (date and time scope of the report)

DATA PREPARATION

WROCŁAWSKI INSTYTUT ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI

PRELIMINARY DATA ANALYSIS

- Tokenizer
- Additional data: female and male names and surnames, postal codes with associated cities and street names
- Morfeusz2

LABELS

- street_start, street_continue, street_no
- company_start, company_continue
- drawing_date_day, drawing_date_month, drawing_date_year
  - dd.mm.yyyy
  - dd month yyyy
- period_from_day, period_from_month, period_from_year
- period_to_day, period_to_month, period_to_year
- postal_code_pre, postal_code_post
- city_start, city_continue, city
- human_start, human_continue, position_start, position_continue

MORFEUSZ2

- Morfeusz2 – an inflectional analyzer and generator for Polish language morphology
- Enables extraction of the basic forms of words (lemmas) together with their features, e.g. tag, commonness, etc.

Range	Segment	Lemma	Tag	Commonness	Qualifiers	Probability	End of sentence
0–1	Dopuszcza	dopuszczać	fin:sg:ter:imperf			1.0000	
1–2	się	się	part			1.0000	
2–3	lokalizację	lokalizacja	subst:sg:acc:f	nazwa_pospolita		1.0000	
3–4	nieuciążliwych	nieuciążliwy	adj:pl:gen:n:pos			1.0000	
4–5	urządzeń	urządzenie	subst:pl:gen:n:ncol	nazwa_pospolita		1.0000	
5–6	komunalnej	komunalny	adj:sg:gen:f:pos			1.0000	
6–7	infrastruktury	infrastruktura	subst:sg:gen:f	nazwa_pospolita		1.0000	
7–8	technicznej	techniczny	adj:sg:gen:f:pos			1.0000	
8–9	.	.	interp			1.0000	●

LABELS

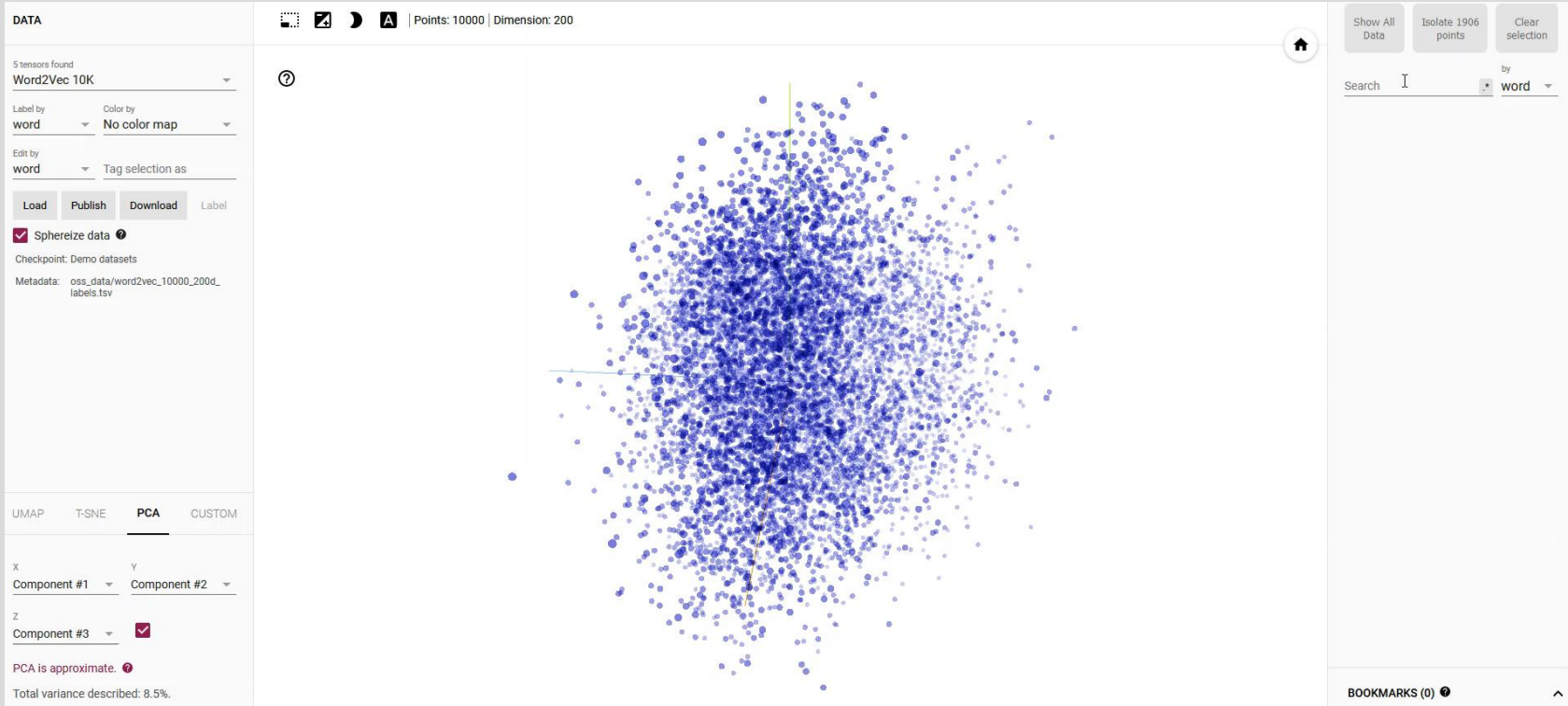
The original text in English	Tokens (original text in Polish)	Label	Lemmata (from Morfeusz2)	Label	What should be extracted
Marie Skłodowska - Curie Street	ul Marii Skłodowskiej Curie	o street_start street_continue street_continue	ul Maria Skłodowska Curie	o o o o	street: ul. Marii Skłodowskiej Curie
Bank Ochrony Środowiska S.A.	Bank Ochrony Środowiska S.A.	company_start company_continue company_continue o	Banek Ochrona Środowisko S.A.	o o o o	company: Bank Ochrony Środowiska S.A.
... together with the president of the board Jerzy Wiśniewski ...	wraz z Prezesem Zarządu Jerzym Wiśniewskim	o o o o o o	wraz z Prezes Zarządu Jerzy Wiśniewski	o o position_start position_continue human_start human_continue	people: Jerzy Wiśniewski Prezes Zarządu

METHODS

WROCŁAWSKI INSTYTUT ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI

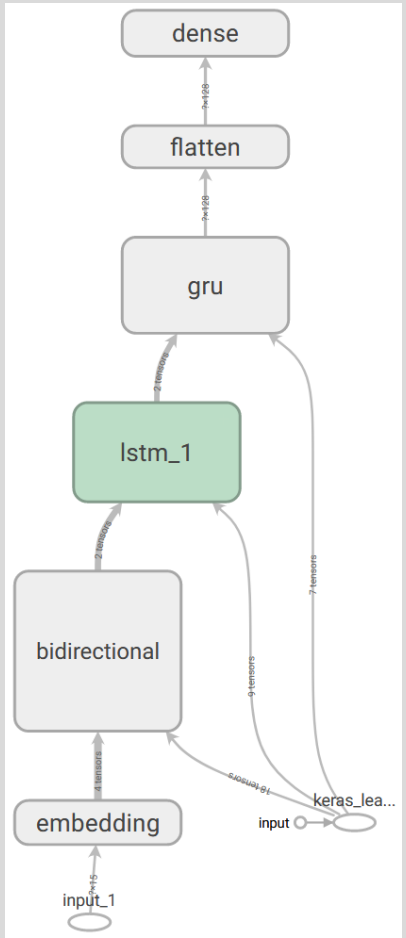
EMBEDDING – PART OF ANN

- Embedding – Layer of each of the considered artificial neural network's structure
- Each word has an assigned *n*-dimensional vector of real numbers, that place it in *n*-dimensional space, thus representing an abstract distance between the words

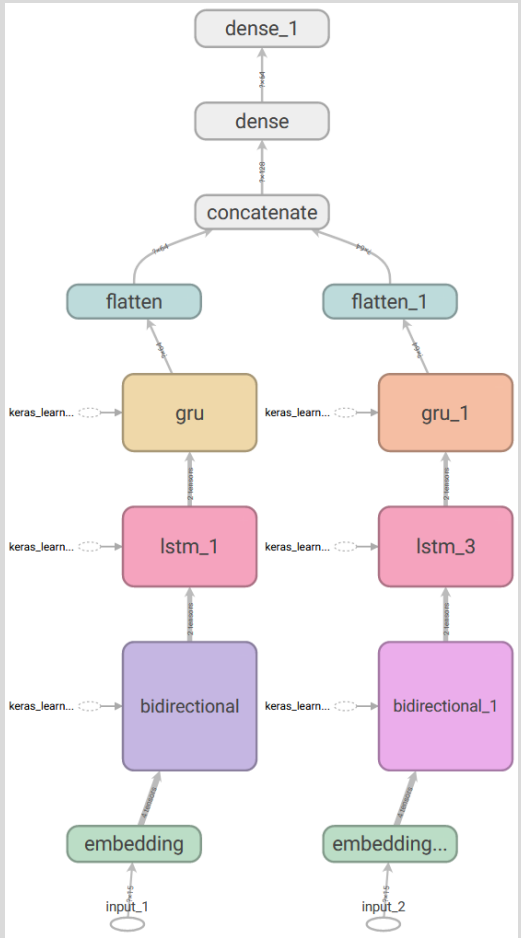


ARTIFICIAL NEURAL NETWORKS

- Both structures were built from the Embedding as the first layer
- Middle – structure with one thread, that as an input takes sequences of words in their original form (tokens)
- Double – structure with two threads, takes as an input two vectors: sequences of tokens and sequences of words in their basic forms (basic)



middle



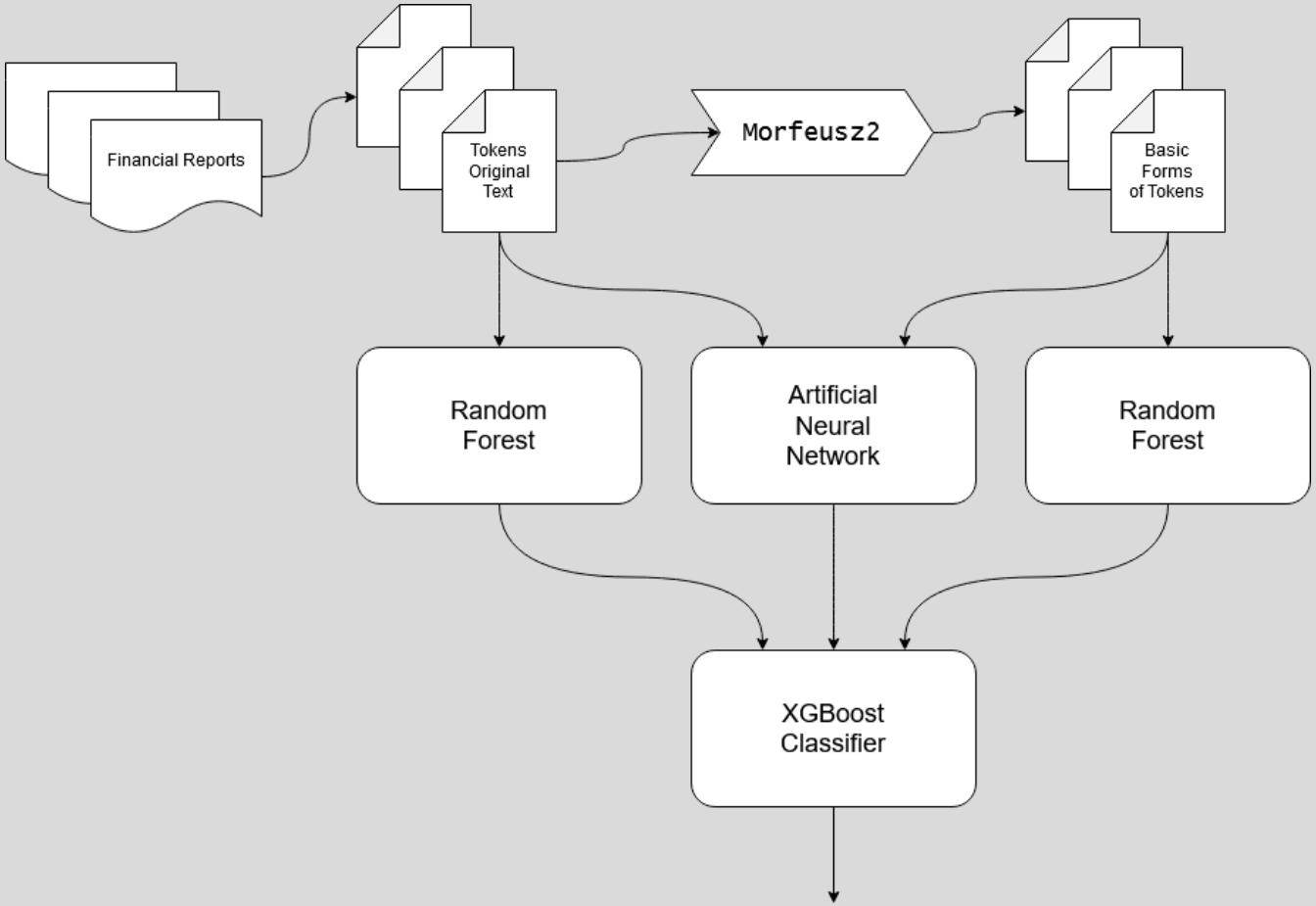
double

RANDOM FOREST AND XGBOOST

- Heterogeneous ensemble consists of members having different base learning algorithms such as SVM, ANN and Decision Trees
- Random Forest – one of the weak learners
- XGBoost – a meta-model, trained to output predictions based on weak learners' predictions

input version	n_estimators	oob_score	prediction score
Token	100	0.9551	0.9959
	200	0.9555	0.9559
	500	0.9956	0.9558
basic	100	0.9586	0.9598
	200	0.9594	0.96
	500	0.9597	0.96

ENSEMBLE MODEL



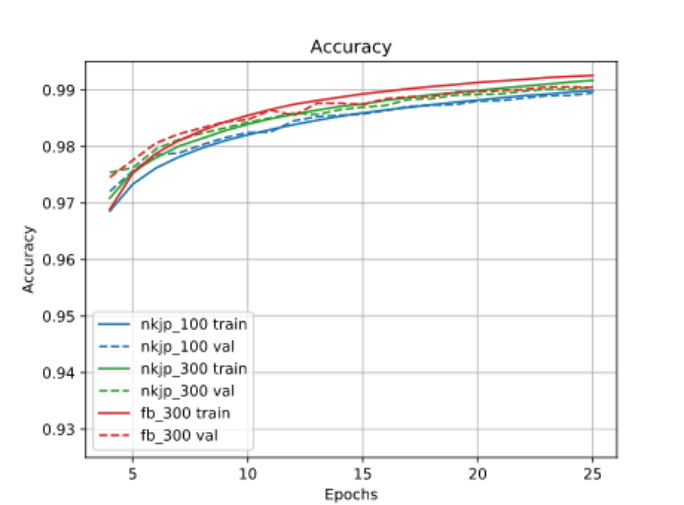
Outputs

Each word in original text has an assigned type (class), e.g. name, surname, company, etc.

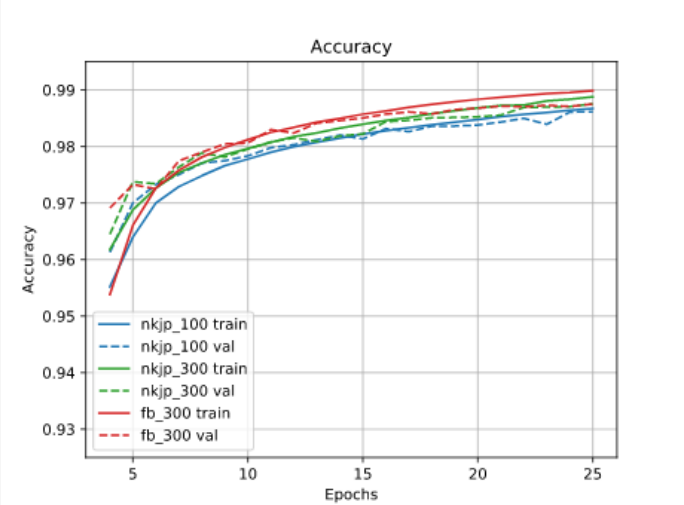
RESULTS

WROCŁAWSKI INSTYTUT ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI

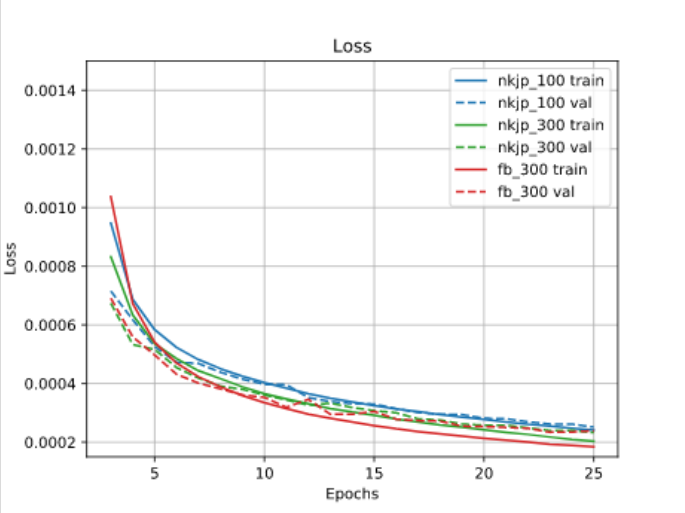
ANNS METRICS



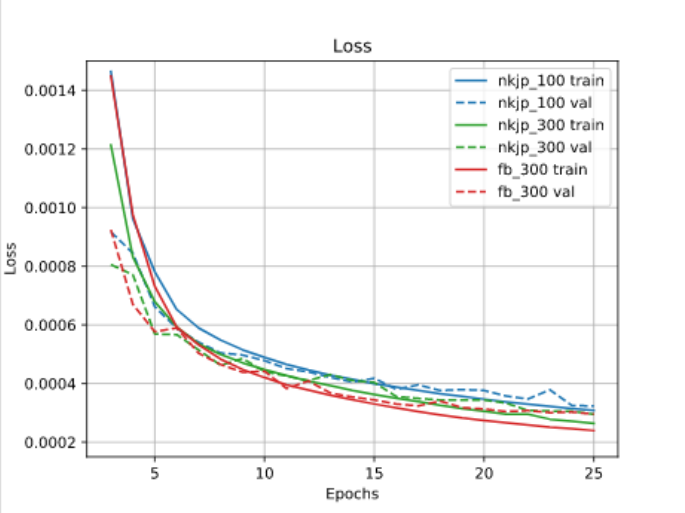
A



B



C



D

INSTYTUT WROCŁAWSKI ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI

F_1 SCORE

Entity	model						
	middle 100 nkjp	middle 300 nkjp	double 100 nkjp	double 300 nkjp	RF 100 nkjp	RF 300 nkjp	XGBoost 300 fb
Company	0.5982	0.6054	0.4126	0.4432	0.6000	0.5784	0.6090
drawing date	0.4667	0.4631	0.4739	0.4667	0.4687	0.4739	0.4767
period from	0.8991	0.9027	0.8883	0.8865	0.8162	0.8054	0.8667
period to	0.9838	0.9784	0.9802	0.9838	0.8378	0.8378	0.9495
postal code	0.6216	0.6216	0.6450	0.6486	0.6378	0.6288	0.6432
City	0.7712	0.7585	0.7748	0.7982	0.7495	0.7441	0.7531
Street	0.4955	0.4793	0.4964	0.4983	0.5063	0.5009	0.5063
street no	0.6252	0.6162	0.6450	0.6324	0.6594	0.6775	0.6793
people names	0.7132	0.6975	0.6539	0.6500	0.7513	0.7585	0.7430
people positions	0.8059	0.8020	0.8139	0.8182	0.8292	0.8396	0.8349

SUMMARY

WROCŁAWSKI INSTYTUT ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI

SUMMARY

model configuration	F_1 score
double_big	0.606 ± 0.017
300_xgb	0.592 ± 0.015
double_small	0.588 ± 0.018
300_RF	0.587 ± 0.015
middle_big	0.585 ± 0.016
100_RF	0.584 ± 0.016

THANK YOU!

WROCŁAWSKI INSTYTUT ZASTOSOWAŃ INFORMACJI PRZESTRZENNEJ I SZTUCZNEJ INTELIGENCJI