“ALEXANDRU IOAN CUZA” UNIVERSITATY OF IAŞI
FACULTY OF COMPUTER SCIENCE

The Semantics and Pragmatics of Natural Language

Daniela GÎFU
http://profs.info.uaic.ro/~daniela.gifu/

Course 7

PREDICTION ALGORITHMS - ECONOMIC CRISIS

SECTION I	SECTION II	SECTION III	SECTION IV	SECTION V
INTRODUCTION	BACKGROUND	DATASET & METHOD	RESULTS & INTERPRETATION	CONCLUSION

The problem

Can a state prevent an economic crisis?

The objectives of this paper:

- to develop a corpus of economic news (in Romanian)
- to provide a proof that analyzing online news can be an effective method for crisis prediction
- to predict a potential economic crisis in Romania

BACKGROUND

The literature

Learning from others…

For prediction of economic crisis:

- discriminate analysis_business failure (Altman, 1968; Beaver, 1966)
- linear conditional probability models (LPM) – (Meyer and Pifer, 1970)
- logistic regression model (Ecer, 2013; Li, Crook, & Andreeva, 2017)
- back propagation neural network (Bell *et al.*, 1990; Brockett *et al*., 1994)
- after 1990 – ML and DL (e.g. decision trees)

DATASET & METHOD

THE CORPUS

RoNews
(Romanian News)

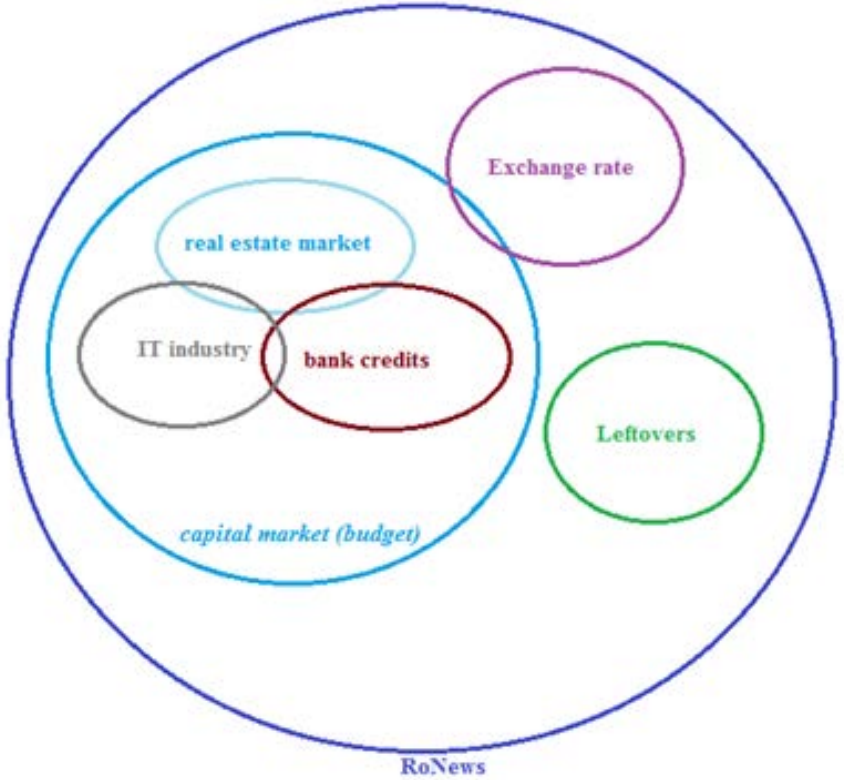
- is a set of news/articles chronologically ordered from 2008 to 2018
- collected from the online publications using an external collector server
- consists of more than 9.000 articles (news)

Publication name	Articles (%)
adevarul.ro	15
agerpres.ro	3
curentul.ro	3
estnews.ro	2
evz.ro	4
gandul.info	15
jurnalul.ro	10
mediafax.ro	4
monitorul.com.ro	5
news.ro	4
romaniatv.ro	4
vremeanoua.ro	4
ziaruldeiasi.ro	2
zf.ro (Ziarul Financiar)	15
others	10

Corpus Categorization

RoNews - 3 categories:

- Exchange rate market
- Capital market (budget)
  - Bank credits
  - Real estate market
  - IT industry
- Leftovers

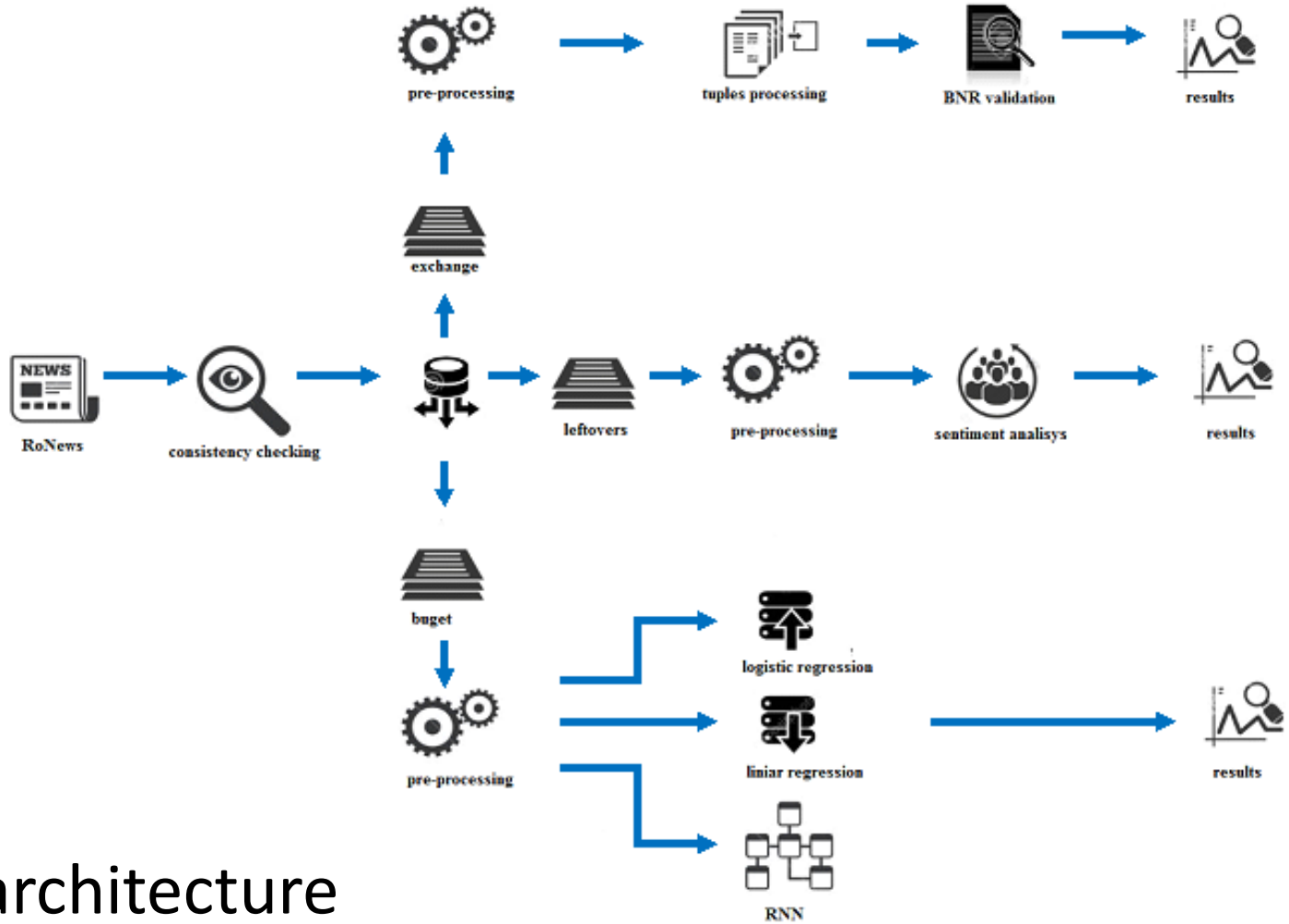


THE METHOD

economicRoTool

4 main modules

- monitoring and data storage from the online news publications & validating articles by using a set of consistency checking algorithms.
- divising corpus into categories and the preprocessing of the articles according to the category to which it belongs.
- processing module, having 3 components for each corpus category.
- results validation and economic crisis prediction.



The architecture

economicPROTOCOL

Dataset preprocessing

- **Consistency checking**
- **Text preprocessing**
  - segmentation
  - tokenization
  - lemmatization
  - POS annotation
  - ……

Consistency check

- validating the authenticity of the RoNews corpus texts

- eliminating duplicates and inconsistent texts of any type

Tool: IBM FileNet Consistency Checker

Results:

the initial corpus size was reduced **9,000** items → **8,400** items

the most affected areas were the **exchange rate market** and **real estate market**

Text preprocessing

- obtaining the same file format for all text in the corpus

- removing links and all unnecessary symbols

- paringe with Python NLTK POS-tagger named TreeTagger

- extracting all items that could indicate calendar dates and amounts of money (RON, EUR, USD).

Processing module

RoNews - 3 categories:

- Exchange rate market

- extract all tuples

`<date, euro value, dollar value>`

form the exchange rate dataset

- applying **BNR validation** algorithm
- graphic view of the results

Example of tuples

Date	EUR	USD
1/15/2013	4.3894	3.2882
1/16/2013	4.3364	3.2596
1/25/2013	4.3606	3.2482
1/28/2013	4.3877	3.2625

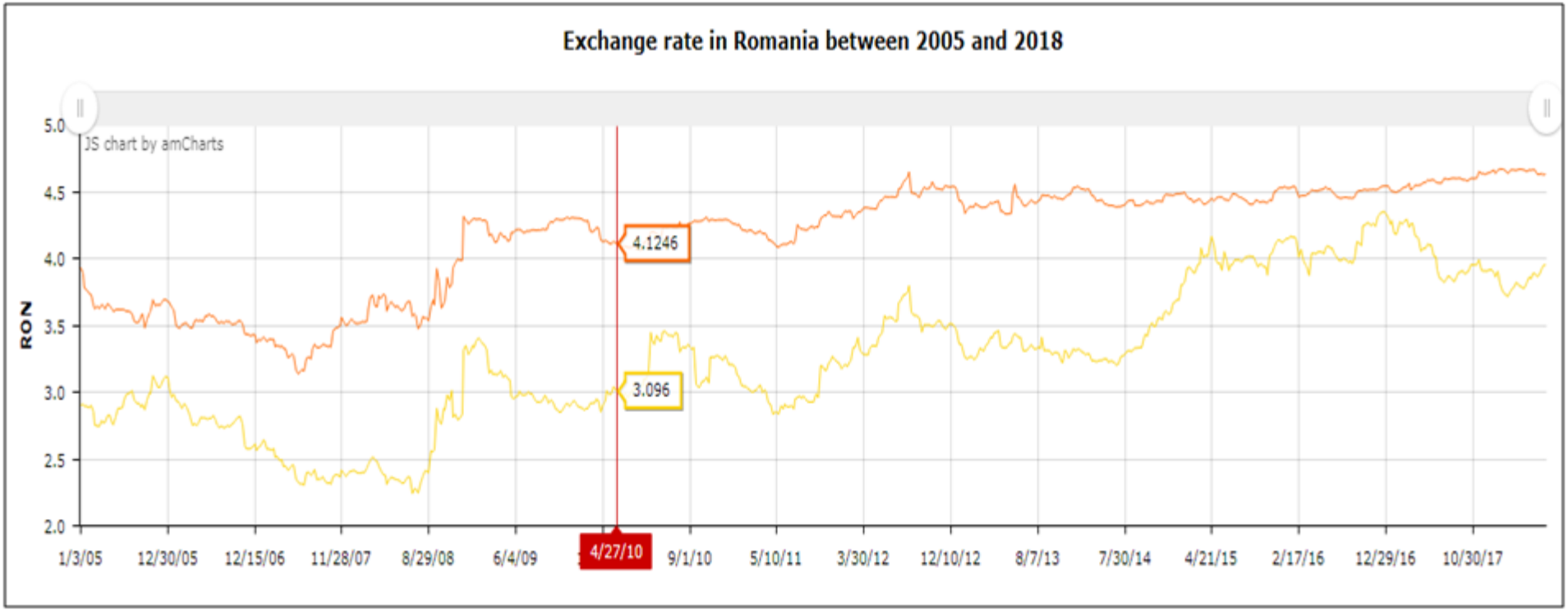
BNR validation

the extracts contained the correct values of the EUR and USD exchange rate from that date

Implementation: a Python script to compare the values with those published by the National Bank of Romania (BNR)

BNR publishes monthly an XML file with all the information on the daily money exchange over the last 13 years

Results: deleting about 20 tuples from the initial set


Exchange rate in Romania between 2005 and 2018
JS chart by amCharts
RON
5.0
4.5
4.0
3.5
3.0
2.5
2.0
4.1246
3.096
1/3/05
12/30/05
12/15/06
11/28/07
8/29/08
6/4/09
4/27/10
9/1/10
5/10/11
3/30/12
12/10/12
8/7/13
7/30/14
4/21/15
2/17/16
12/29/16
10/30/17

Processing module

RoNews - 3 categories:

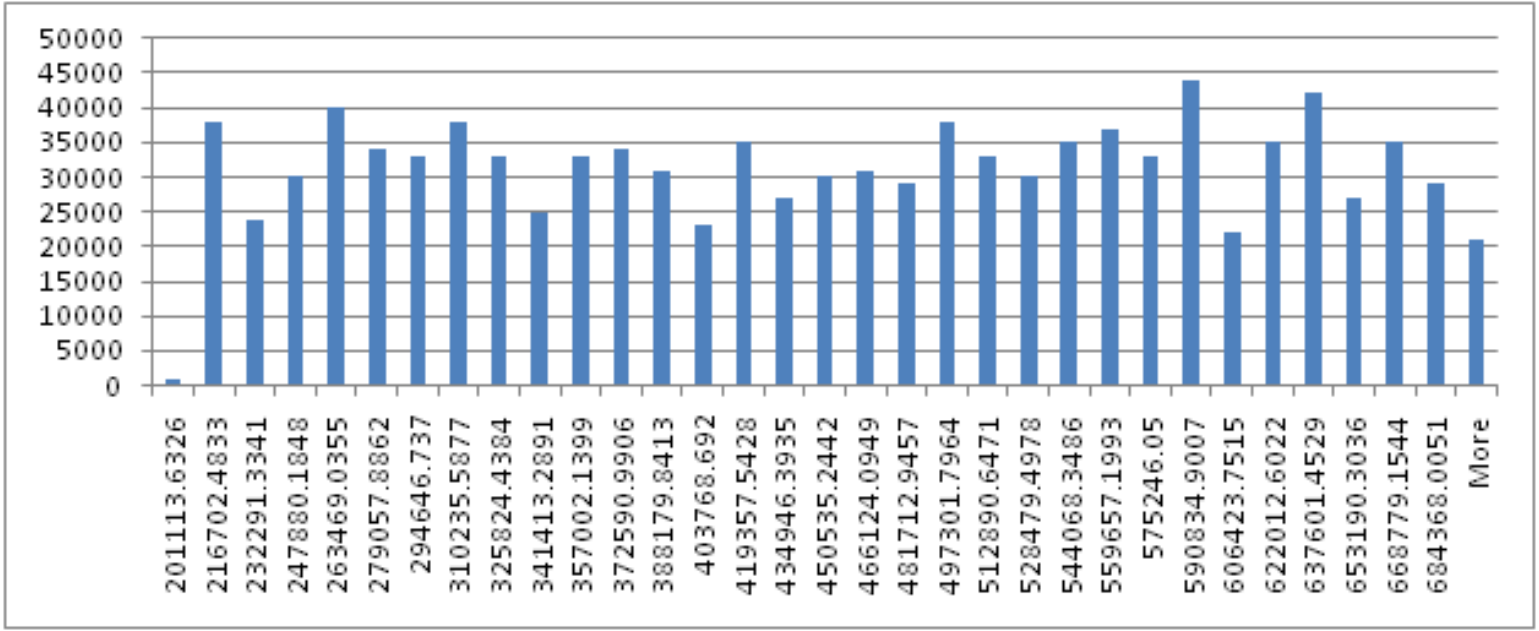
- Capital Market

- **Bank credits**
  - Logistic Regression
- **Real estate market**
  - Linear Regression
- **IT industry**
  - RNN (Recurrent Neural Network)

Logistic Regression algorithm

- dataset ≈ 2000 files dealing with bank credits
- only 1030 files could be extracted accurately

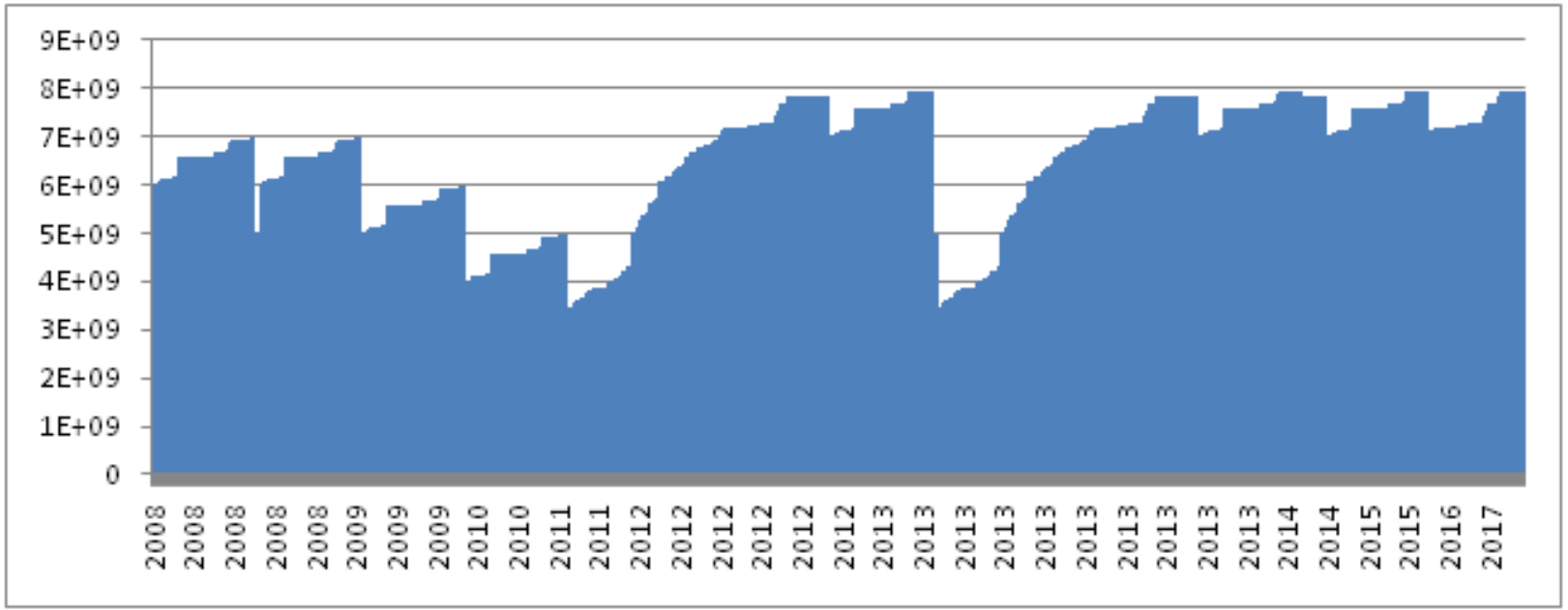
Results: the ratio between the number of bank credits and their values



Linear Regression algorithm

- dataset ≈ 1000 files dealing with the real estate market
- only 611 files could be extracted accurately

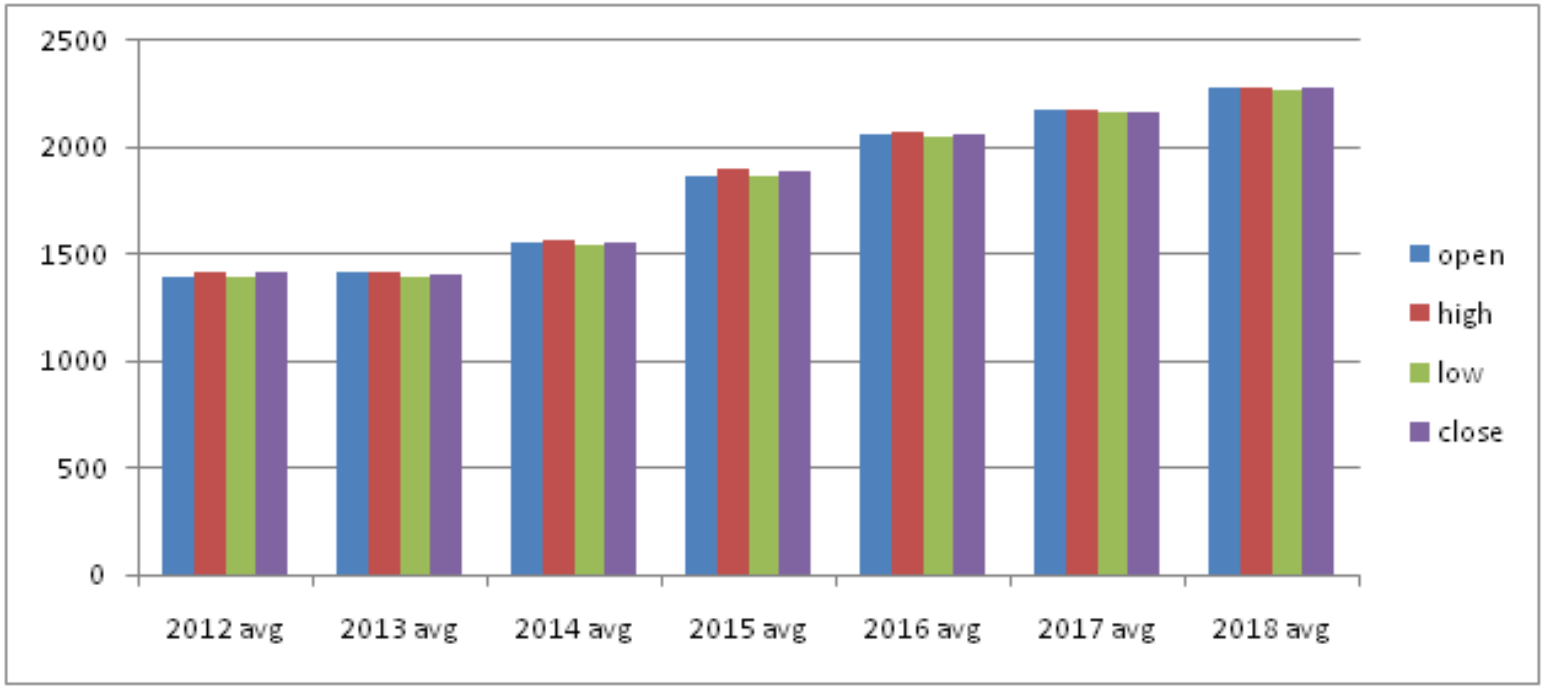
Results: the route followed by the real estate market from 2008 to the present



RNN (Recurrent Neural Network)

- training dataset ≈ 1200 news, containing only text related with IT industry

Results: evolution of the IT industry from 2012 to 2018



Processing module

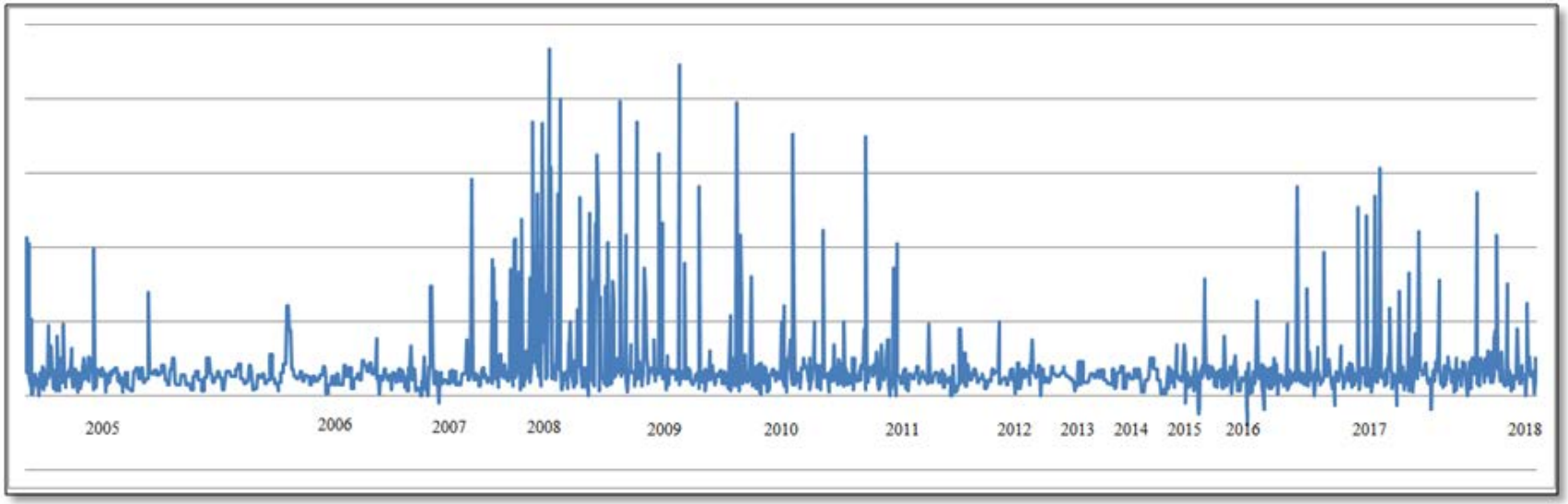
RoNews - 3 categories:

- Leftovers

- a Sentiment Analysis (SA) method
- tracing sentiment classes (negative and positive) dominated by those news

Sentiment Analysis

- dataset ≈ 1900 files
- using a set of Python libraries - Polyglot



RESULTS & INTERPRETATIONS

Dataset - Statistics

Total number of news (approximate)	Inconsistent or duplicates (approximate)	Final number of news (approximate)
9.000	600	8.400

The corpus categorization

Module name	Algorithm (method) name	Number of texts (approximate)	Number of accurate texts
Exchange rate	BNR validation	2100	1900
Bank credits	Logistic Regression	2000	1030
Real estate market	Linear Regression	1000	611
IT industry	RNN	1200	1200
Leftovers	Sentiment Analysis	1900	1900

Results

Exchange rate processing	Accuracy	0.779
	Recall	0.797
	Precision	0.765

Logistic Regression	Accuracy	0.742
	Recall	0.733
	Precision	0.766

Linear Regression	Accuracy	0.784
	Recall	0.798
	Precision	0.765

RNN	Accuracy	0.731
	Recall	0.754
	Precision	0.734

Sentiment Analysis	Accuracy	0.854
	Recall	0.812
	Precision	0.845

CONCLUSIONS

- developing a corpus of economic news in Romanian
- providing a proof that analyzing online news using NLP concepts - an effective method for prediction of an economic crisis (Romanian case)
- predicting the next crisis in Romania (around 2020)

Is it our prediction accurate enough with this corpus?

economicROTOOL has an error margin of approximately one or two years that the signs of the next economic crisis will be observed over **2 years.**

Here it is a probability about the next crisis - **officially declared over 5 years** (around 2023???)

.

Thank you for your attention!