ERSTE GROUP

Risk-Relevant Information in Czech Newspaper Articles: Simple Text-Mining Model

Jozef Zubrický, Petr Kadeřábek
Erste Group Bank
24.03.2016

There is lot of unstructured data information in Czech Newspapers regarding companies

ERSTE GROUP

Web ČTK | Infobanka | Fotobanka | Videobanka | Infografika | Akademie ČTK | PR servis | České noviny

ČTK ČESKÉ NOVINY

ČR | Svět | Ekonomika | Kultura

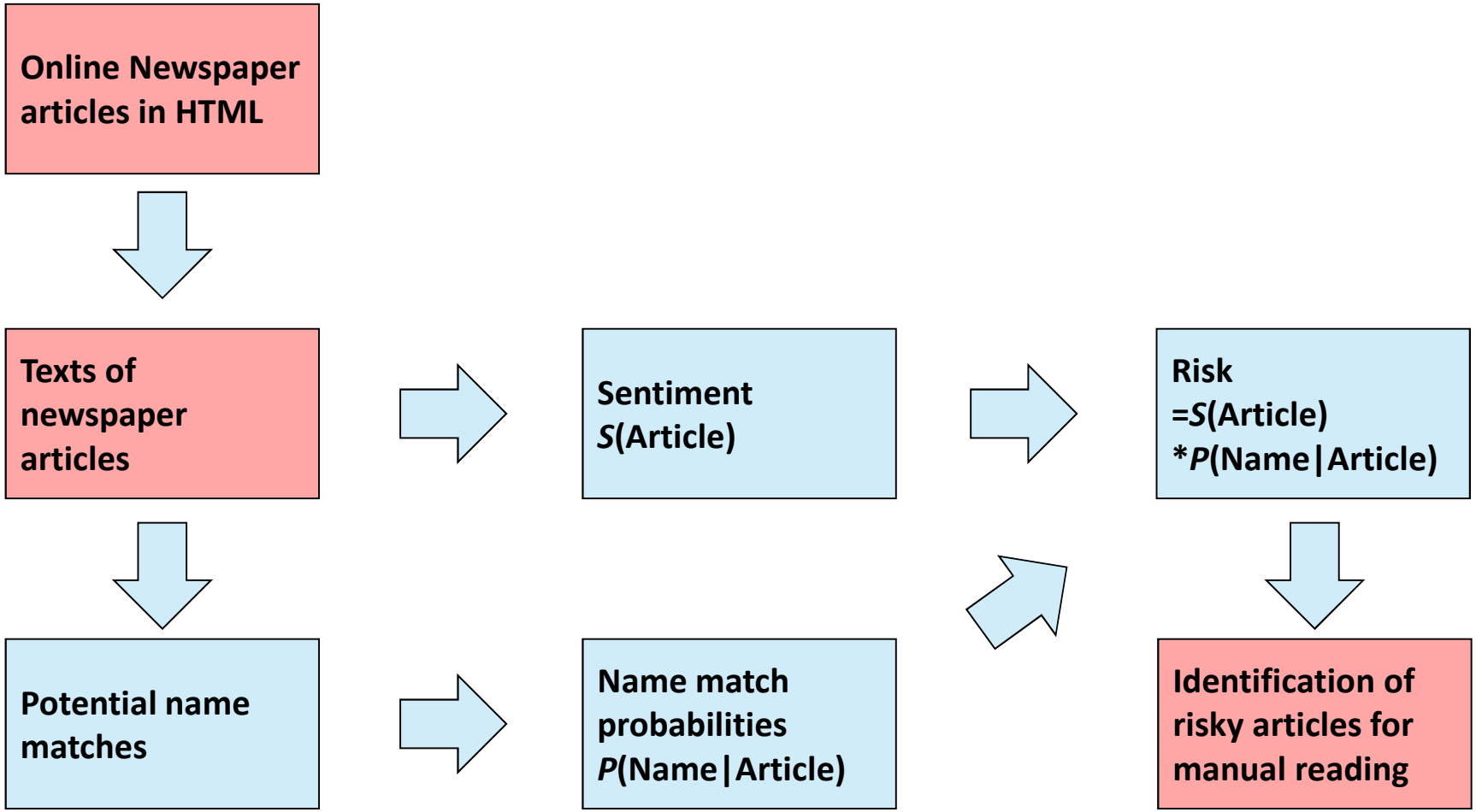
Amsterdam - Leading Czech entrepreneur and investor Zdenek Bakala has not been elected onto the board of directors of mining company New World Resources (NWR) at today´s general meeting and former investment banker Charles Harman took his place, according to information from NWR´s website.

Bakala said in the middle of March already that he would leave NWR´s management at the general meeting. NWR is the owner of black coal mining company OKD.

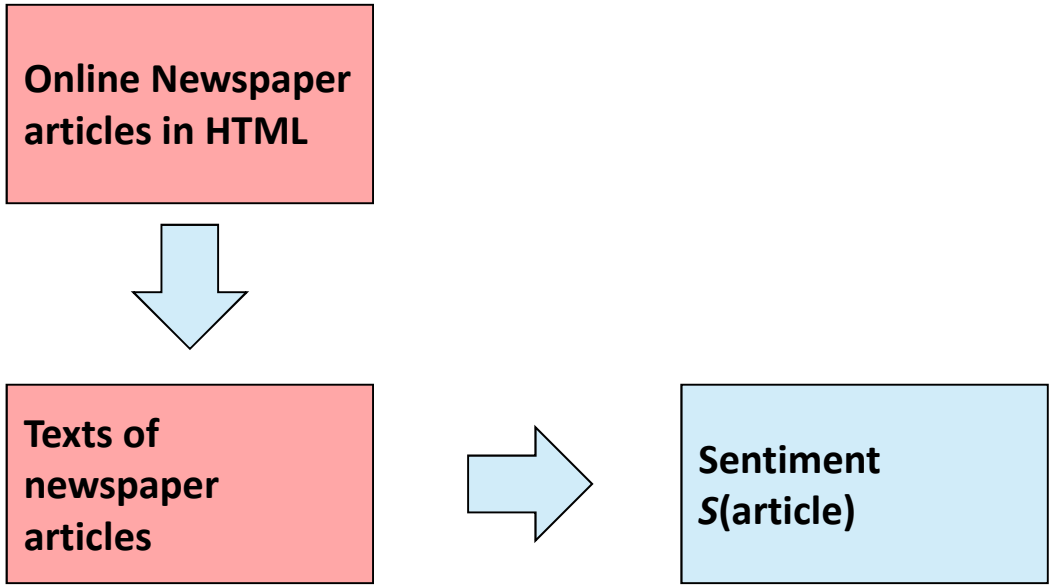
Bakala, one of the richest Czechs, acquired a stake in OKD in 2004 and became a co-owner in NWR, a holding company for mining assets, based in the Netherlands.

After the completion of NWR´s financial restructuring last year, Bakala owns 50 percent in group CERCL Mining, former BXR mining, which holds 50.5 percent in NWR.

Our goal was to create a simple but performing model utilizing this information in our credit decisions and early warning systems



One problem is to calcualate the sentiment of the article



First step is to divide document into sets of words with the same meaning

ERSTE GROUP

Stemming/Lemmatization:

- Grouping occurrences of different grammatical forms of the same word
- Stemma of „meeting“ is „meet“ Lemma of „better“ is „good“
- Stemming is simpler, faster but works on word by word basis without context of the works
- Lemmatization is more difficult but ensures „lazy“ and „laziness“ is lemmatized to the same word
- Available in SAS Text Miner but free tools also available (Czech stemmer in Snowball, Majka)

Next Step is to create so called „terms“ of connected words that produce meaning

ERSTE GROUP

„Term” denotes 1,2,3-gram of stems or lemmas

- n-gram consists of n consecutive words (stems, lemmas)
- Higher n-gram means higher computation complexity
- Used by Google for next suggested word
- Actual level of n-gram may be set in SAS Text Miner

	1-gram	2-gram	3-gram
Bankruptcy	Bankrupt	Bankruptcy occurred	Bankruptcy occurred on
		Bankruptcies happened	Bankruptcies occurring at
			Bankruptcy occurs in

Next Step is to create the „terms“ per document frequency matrix

ERSTE GROUP

Term Matrix can be set up in SAS **Text Miner**

	Doc1	Doc2	Doc3	Doc4	Doc5
Bankrupt	3	2	1	0	0
Insolvency	1	2	2	0	0
Technology	0	0	0	1	2
Success	0	0	0	0	2

Due to lack of defaults and to avoid overfitting we used the Latent Semantic Indexing method to estimate sentiment from the term-frequency matrix.

ERSTE GROUP

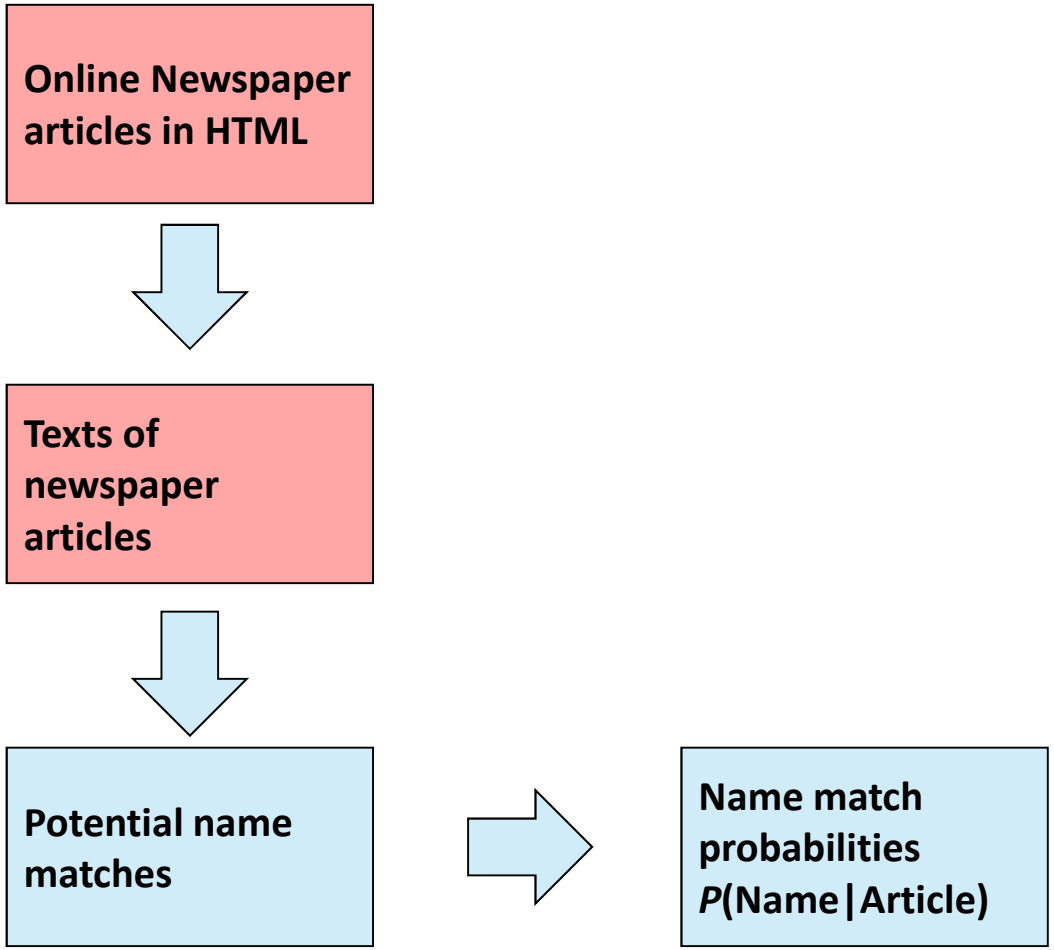
Latent Semantic Indexing (LSI):

- Algebraically simple method (based on Singular Value Decomposition)
- It de-composes „term“ matrix into matrix enriched by words used in the same concept as sentiment words e.g. „bankrupt“ and „layoffs“
- You start with typically negative sentiment words, e.g. „bankrupt“ or „insolvency“
- LSI identifies also the articles containing the word “layoffs”, although they do not contain “bankrupt” directly.

	Doc1	Doc2	Doc3	Doc4	Doc5
Bankrupt	2.9	1.9	0.9	0.1	0.1

Word „Layoff“ no directly „Bankrupt“

Second problem is to make find articles that are written about the company you need



The probabilistic approach to named-entity recognition was chosen

ERSTE GROUP

For Example Apple, Inc. Is a company but apple is used in lot of other contexts than speaking about computers

- Our rules basic rules:
  - 1st character not following . is an upper-case letter or a digit
  - Ignore names identical to very common words e.g. apple ☺

Name match validity is affected by many factors

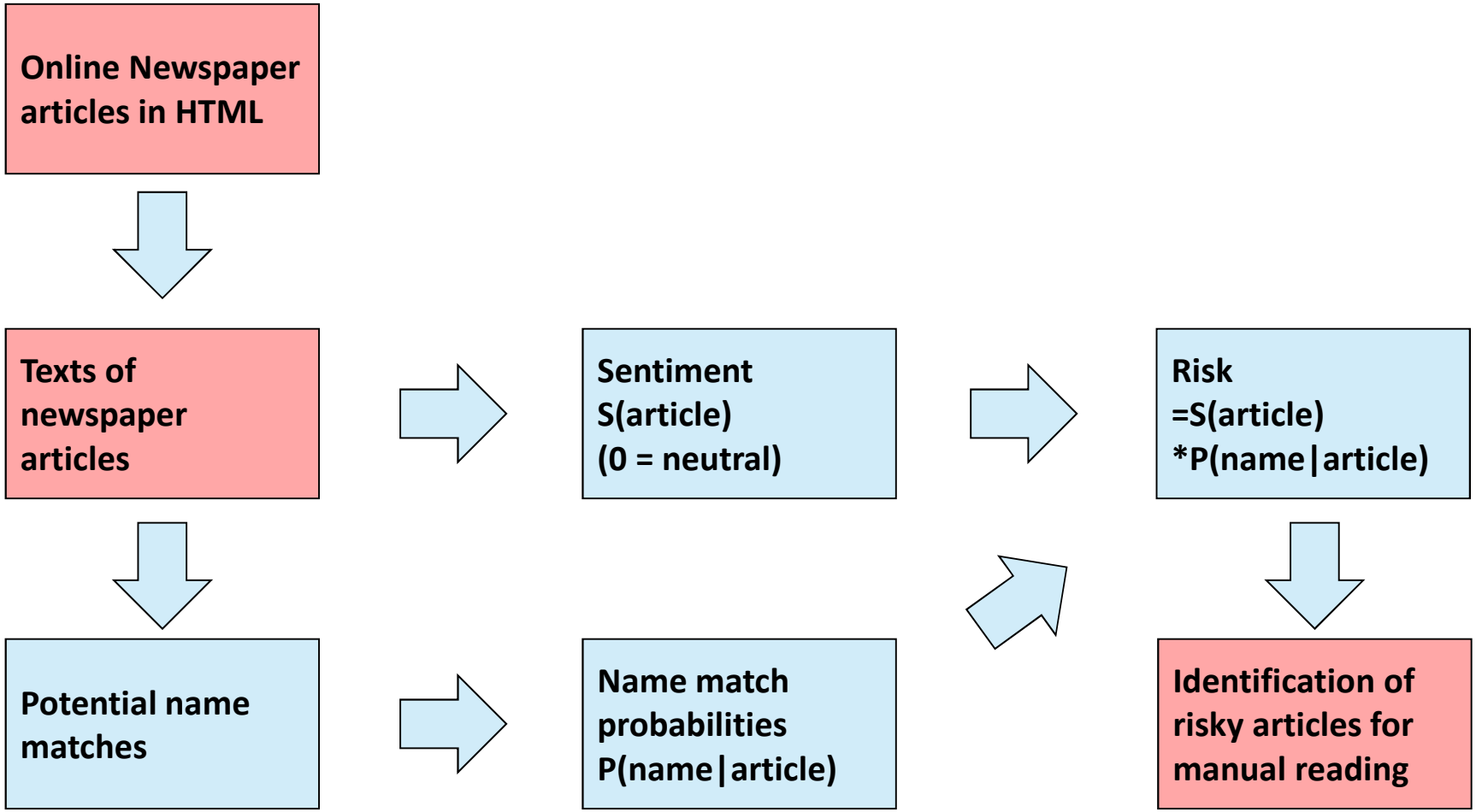
ERSTE GROUP

Step 2: Logistic regression model trained on the manually-tagged potential matches

- **Target: valid/invalid manually checked match**
- **Main explanatory variables contributing to higher match probability:**
  - **Word like "company" before the potential name**
  - **Legal form after the name e.g. „, s.r.o."**
  - **Length of the name**
  - **Frequency of name occurrence in other contexts (occurrence in articles with lower-case first letter, occurrence in lists of company names, personal names etc.)**

In the end we need to put this two pieces together

ERSTE GROUP



Use of the results in the model of probability of default

ERSTE GROUP

Currently used simple composition approach

- Sentiment = 0 denotes neutral information regardless of the match validity.
- Match valid with 0% probability has neutral information regardless of the article sentiment.
- The following variable *X* will be used in the Probability of default model:

$$X = S(\text{Article}) * P(\text{Name} | \text{Article})$$

Use of the model in the bank

ERSTE GROUP

Short term

- Select articles showing highest and lowest risk for manual reading.
- Distribute only the relevant articles based on sentiment to relevant relationship managers based on the name match
- High-risk sentiment articles used by risk management.
- Low-risk sentiment articles indicate a business opportunity.

Mid term

- Use of the information from newspaper articles directly in the rating models.