Assessing Causality Structures Learned from Digital Text Media





Maisonnave Mariano
Universidad Nacional del Sur

mariano.maisonnave@cs.uns.edu.ar

https://cs.uns.edu.ar/~mmaisonnave

Fernando Delbianco
Universidad Nacional del Sur

fernando.delbianco@uns.edu.ar

Fernando A. Tohmé
Universidad Nacional del Sur

ftohme@criba.edu.ar

Ana G. Maguitman
Universidad Nacional del Sur

agm@cs.uns.edu.ar

Evangelos E. Milios
Dalhousie University

eem@cs.dal.ca

Motivation

The Guardian

The New York Times

twitter

yahoo! finance

THE WORLD BANK

Motivation

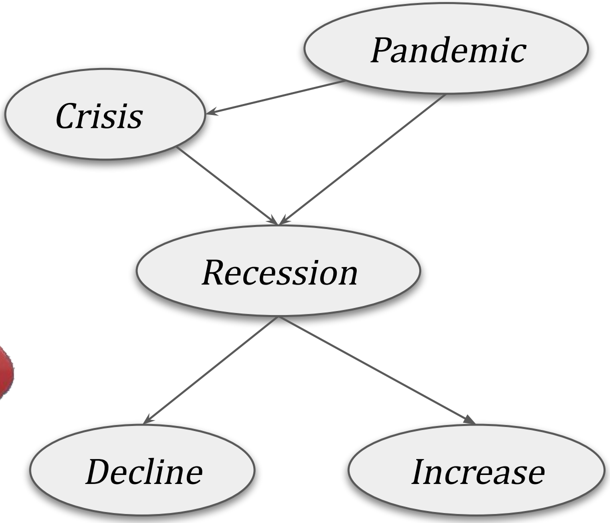
The Guardian

The New York Times

twitter

yahoo! finance

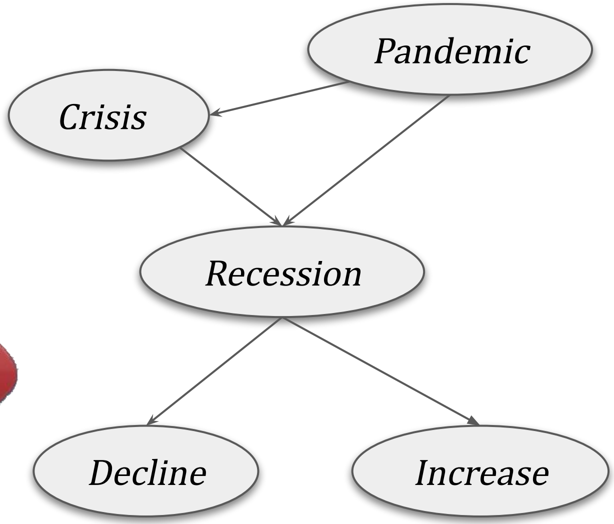
THE WORLD BANK



Motivation

Allow domain experts to

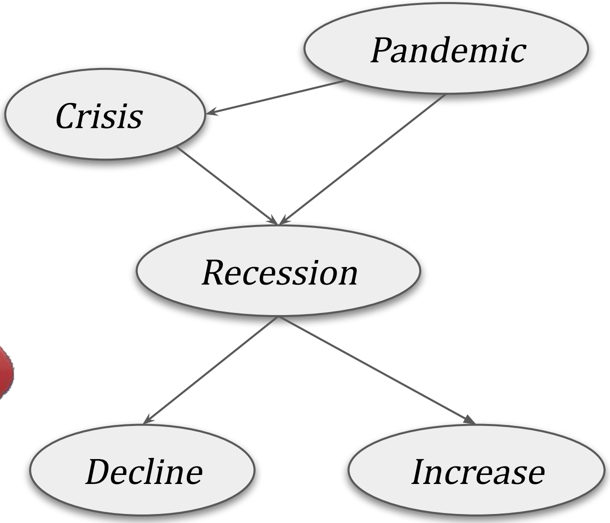
- Understand



Motivation

Allow domain experts to

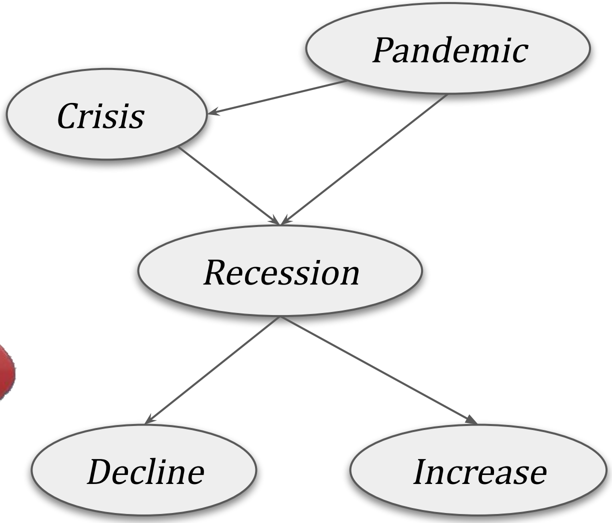
- Understand
- Explain



Motivation

Allow domain experts to

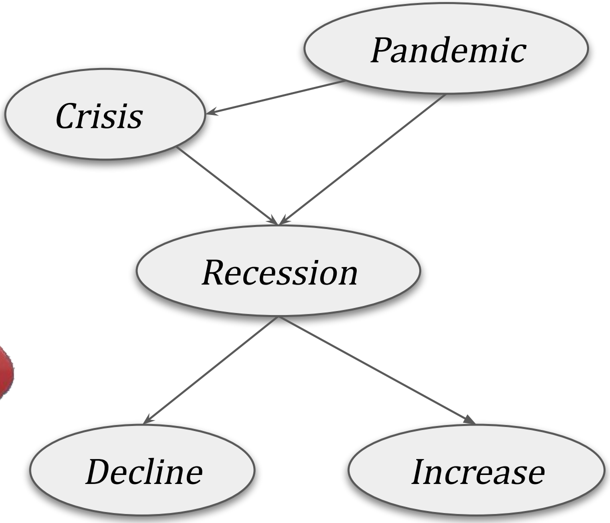
- Understand
- Explain
- Predict



Motivation

Allow domain experts to

- Understand
- Explain
- Predict

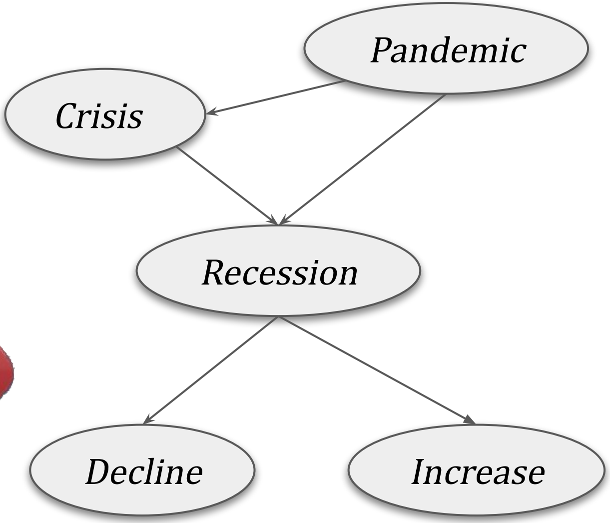


1. What are those variables? How do we extract them?

Motivation

Allow domain experts to

- Understand
- Explain
- Predict

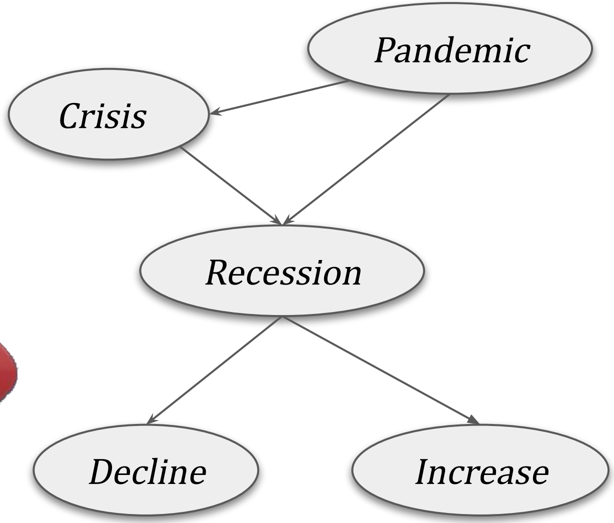


1. What are those variables? How do we extract them?
2. From the extracted variables, which ones are relevant?

Motivation

Allow domain experts to

- Understand
- Explain
- Predict

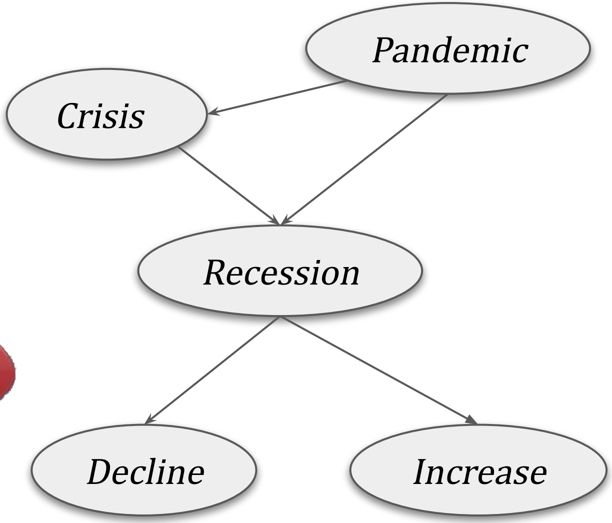


1. What are those variables? How do we extract them?
2. From the extracted variables, which ones are relevant?
3. How do we incorporate other sources of information?

Motivation

Allow domain experts to

- Understand
- Explain
- Predict

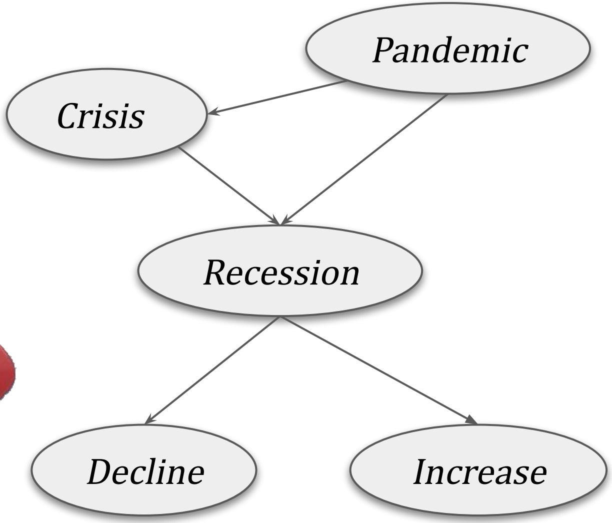


1. What are those variables? How do we extract them?
2. From the extracted variables, which ones are relevant?
3. How do we incorporate other sources of information?
4. How do we build a causal model from all these variables?

Motivation

Allow domain experts to

- Understand
- Explain
- Predict



1. What are those variables? How do we extract them?
2. From the extracted variables, which ones are relevant?
3. How do we incorporate other sources of information?
4. How do we build a causal model from all these variables?
5. How do we display this information effectively?

Proposed Framework

Proposed Framework

1. What are those variables? How do we extract them?
2. From the extracted variables, which ones are relevant?
3. How do we incorporate other sources of information?
4. How do we build a causal model from all these variables?
5. How do we display this information effectively?

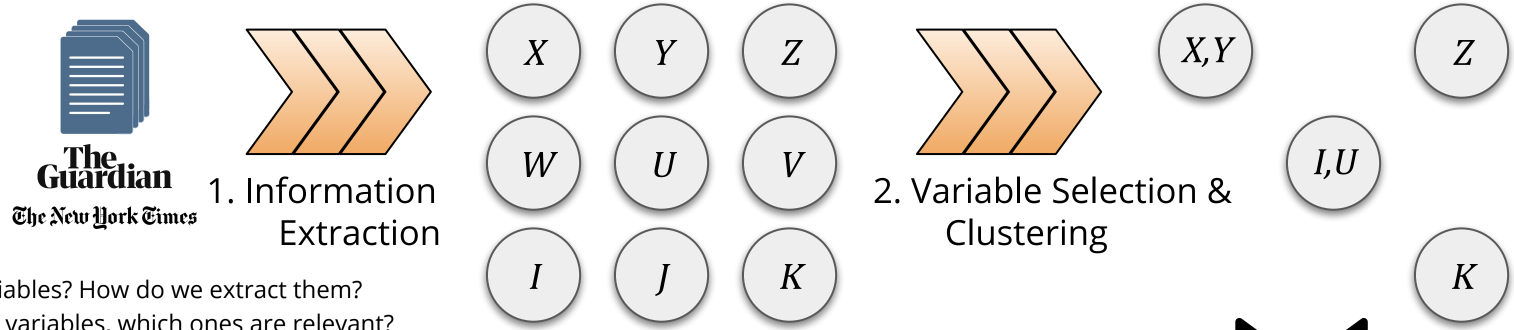
Proposed Framework

1. **What are those variables? How do we extract them?**
2. From the extracted variables, which ones are relevant?
3. How do we incorporate other sources of information?
4. How do we build a causal model from all these variables?
5. How do we display this information effectively?

Proposed Framework

1. What are those variables? How do we extract them?
2. **From the extracted variables, which ones are relevant?**
3. How do we incorporate other sources of information?
4. How do we build a causal model from all these variables?
5. How do we display this information effectively?

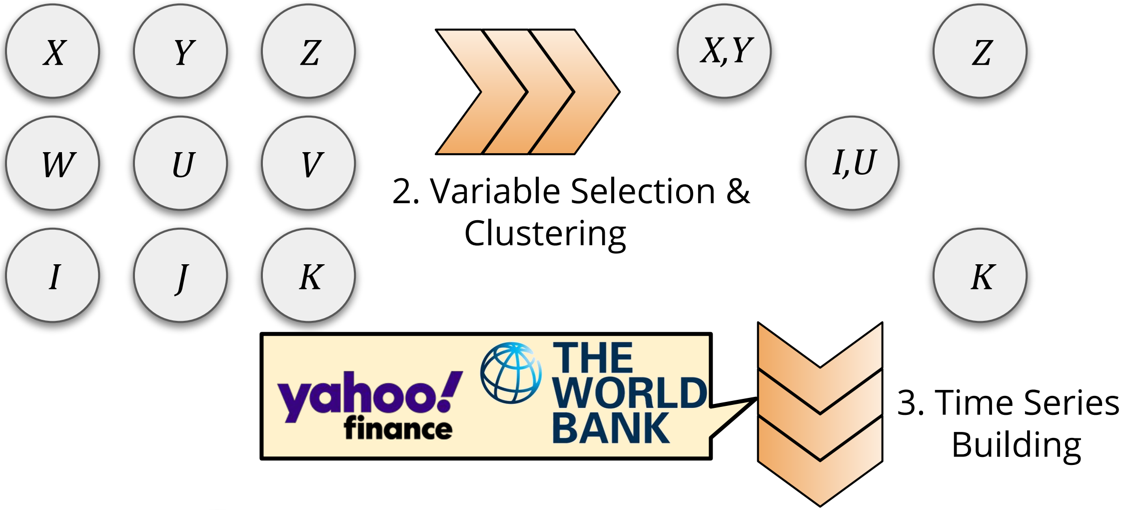
Proposed Framework



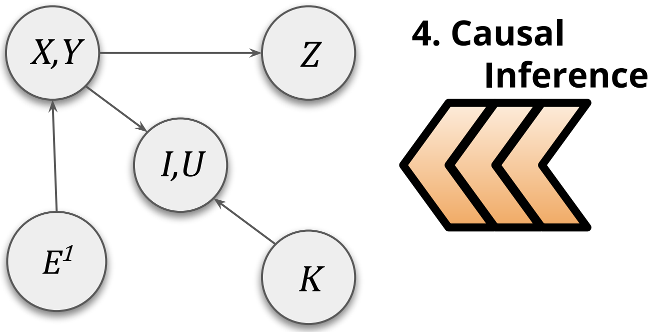
1. What are those variables? How do we extract them?
2. From the extracted variables, which ones are relevant?
3. **How do we incorporate other sources of information?**
4. How do we build a causal model from all these variables?
5. How do we display this information effectively?

t	XY_t	Z_t	IU_t	J_t	$E^1{}_t$
$t=0$	XY_0	Z_0	IU_0	I_0	$E^1{}_0$
$t=1$	XY_1	Z_1	IU_1	I_1	$E^1{}_1$
...	...	...	...	...	...

Proposed Framework

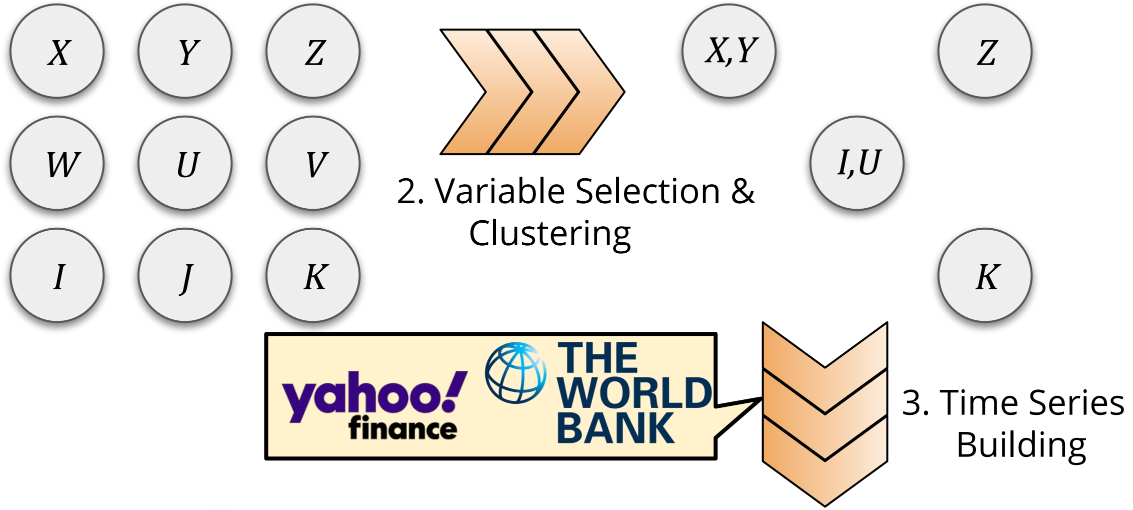


1. What are those variables? How do we extract them?
2. From the extracted variables, which ones are relevant?
3. How do we incorporate other sources of information?
4. **How do we build a causal model from all these variables?**
5. How do we display this information effectively?

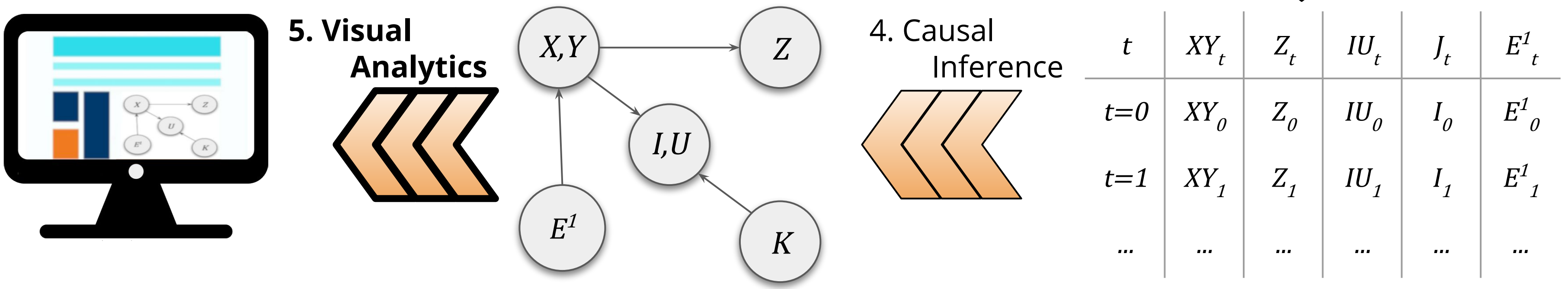


t	XY_t	Z_t	IU_t	J_t	$E^1{}_t$
$t=0$	XY_0	Z_0	IU_0	I_0	$E^1{}_0$
$t=1$	XY_1	Z_1	IU_1	I_1	$E^1{}_1$
...	...	...	...	...	...

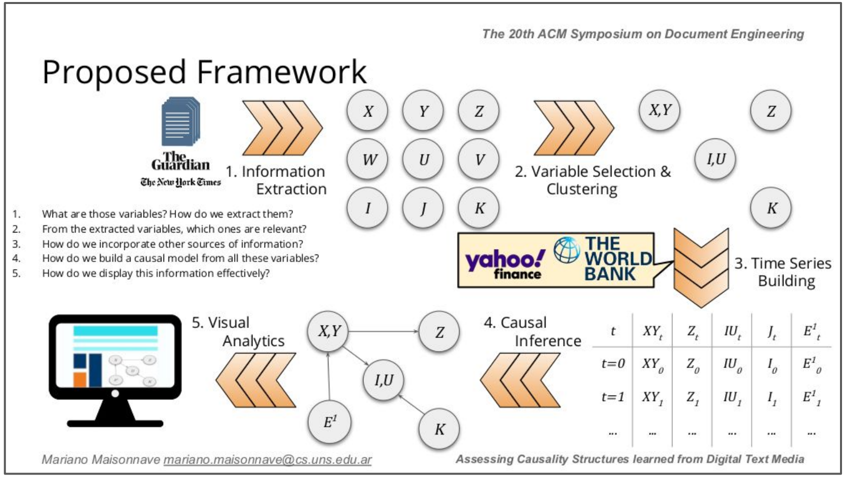
Proposed Framework



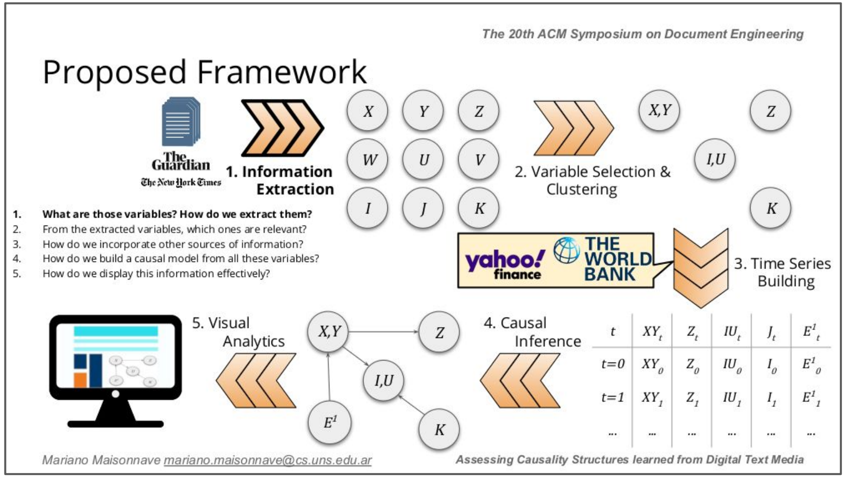
1. What are those variables? How do we extract them?
2. From the extracted variables, which ones are relevant?
3. How do we incorporate other sources of information?
4. How do we build a causal model from all these variables?
5. **How do we display this information effectively?**



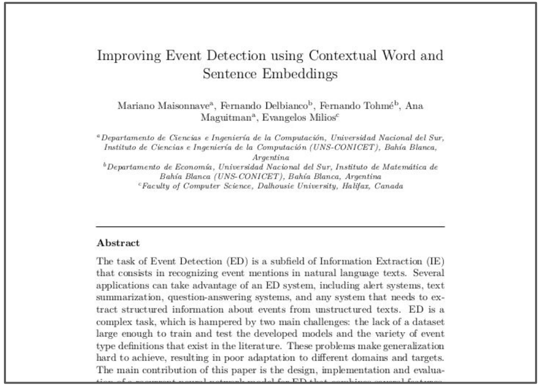
Results


The 20th ACM Symposium on Document Engineering
Proposed Framework
The Guardian
The New York Times
1. Information Extraction
X
Y
Z
W
U
V
I
J
K
2. Variable Selection & Clustering
X,Y
Z
I,U
K
1. What are those variables? How do we extract them?
2. From the extracted variables, which ones are relevant?
3. How do we incorporate other sources of information?
4. How do we build a causal model from all these variables?
5. How do we display this information effectively?
yahoo! finance
THE WORLD BANK
3. Time Series Building
5. Visual Analytics
4. Causal Inference
t | XY_t | Z_t | IU_t | J_t | E^1_t
t=0 | XY_0 | Z_0 | IU_0 | I_0 | E^1_0
t=1 | XY_1 | Z_1 | IU_1 | I_1 | E^1_1
... | ... | ... | ... | ... | ...
E^1
Mariano Maisonnave mariano.maisonnave@cs.uns.edu.ar
Assessing Causality Structures learned from Digital Text Media

Results



- Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A. and Milios, E., 2020. Improving Event Detection using Contextual Word and Sentence Embeddings. *arXiv preprint arXiv:2007.01379*.



Improving Event Detection using Contextual Word and Sentence Embeddings

Mariano Maisonnave[a], Fernando Delbianco[b], Fernando Tohmé[b], Ana Maguitman[a], Evangelos Milios[c]

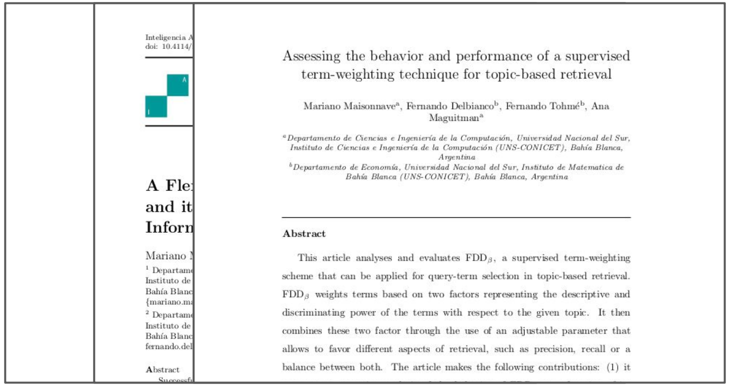
[a]Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Instituto de Ciencias e Ingeniería de la Computación (UNS-CONICET), Bahía Blanca, Argentina
[b]Departamento de Economía, Universidad Nacional del Sur, Instituto de Matemática de Bahía Blanca (UNS-CONICET), Bahía Blanca, Argentina
[c]Faculty of Computer Science, Dalhousie University, Halifax, Canada

Abstract

The task of Event Detection (ED) is a subfield of Information Extraction (IE) that consists in recognizing event mentions in natural language texts. Several applications can take advantage of an ED system, including alert systems, text summarization, question-answering systems, and any system that needs to extract structured information about events from unstructured texts. ED is a complex task, which is hampered by two main challenges: the lack of a dataset large enough to train and test the developed models and the variety of event type definitions that exist in the literature. These problems make generalization hard to achieve, resulting in poor adaptation to different domains and targets. The main contribution of this paper is the design, implementation and evalua-

Results

- Maisonnave, M., Delbianco, F., Tohmé, F.A. and Maguitman, A.G., 2018. A Supervised Term-Weighting Method and its Application to Variable Extraction from Digital Media. In *XIX Simposio Argentino de Inteligencia Artificial (ASAI)-JAIIO 47 (CABA, 2018)*.
- Maisonnave, M., Delbianco, F., Tohmé, F.A. and Maguitman, A.G., 2019. A flexible supervised term-weighting technique and its application to variable extraction and information retrieval. *Inteligencia Artificial*, *22*(63), pp.61-80.
- Maisonnave, M., Delbianco F., Tohmé, F. and Maguitman, A., 2020. A supervised term-weighting technique for topic-based retrieval. *arXiv preprint arXiv:2007.06616*.



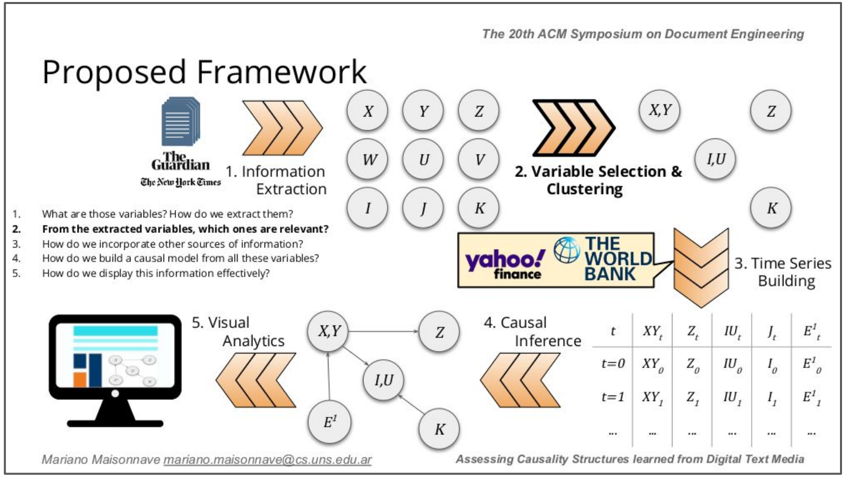
Assessing the behavior and performance of a supervised term-weighting technique for topic-based retrieval

Mariano Maisonnave[a], Fernando Delbianco[b], Fernando Tohmé[b], Ana Maguitman[a]

[a]Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Instituto de Ciencias e Ingeniería de la Computación (UNS-CONICET), Bahía Blanca, Argentina
[b]Departamento de Economía, Universidad Nacional del Sur, Instituto de Matematica de Bahía Blanca (UNS-CONICET), Bahía Blanca, Argentina

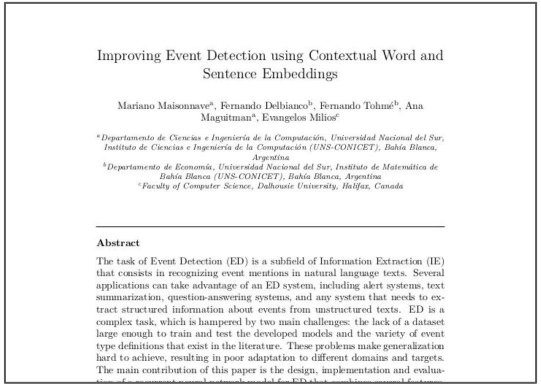
Abstract

This article analyses and evaluates FDD_β, a supervised term-weighting scheme that can be applied for query-term selection in topic-based retrieval. FDD_β weights terms based on two factors representing the descriptive and discriminating power of the terms with respect to the given topic. It then combines these two factor through the use of an adjustable parameter that allows to favor different aspects of retrieval, such as precision, recall or a balance between both. The article makes the following contributions: (1) it



t	XY_t	Z_t	IU_t	J_t	E^1_t
t=0	XY_0	Z_0	IU_0	I_0	E^1_0
t=1	XY_1	Z_1	IU_1	I_1	E^1_1
...	...	...	...	...	...

Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A. and Milios, E., 2020. Improving Event Detection using Contextual Word and Sentence Embeddings. *arXiv preprint arXiv:2007.01379*.



Improving Event Detection using Contextual Word and Sentence Embeddings

Mariano Maisonnave[a], Fernando Delbianco[b], Fernando Tohmé[b], Ana Maguitman[a], Evangelos Milios[c]

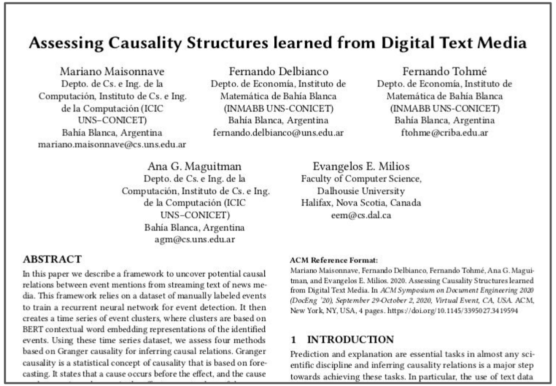
[a]Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Instituto de Ciencias e Ingeniería de la Computación (UNS-CONICET), Bahía Blanca, Argentina
[b]Departamento de Economía, Universidad Nacional del Sur, Instituto de Matemática de Bahía Blanca (UNS-CONICET), Bahía Blanca, Argentina
[c]Faculty of Computer Science, Dalhousie University, Halifax, Canada

Abstract

The task of Event Detection (ED) is a subfield of Information Extraction (IE) that consists in recognizing event mentions in natural language texts. Several applications can take advantage of an ED system, including alert systems, text summarization, question-answering systems, and any system that needs to extract structured information about events from unstructured texts. ED is a complex task, which is hampered by two main challenges: the lack of a dataset large enough to train and test the developed models and the variety of event type definitions that exist in the literature. These problems make generalization hard to achieve, resulting in poor adaptation to different domains and targets. The main contribution of this paper is the design, implementation and evalua-

Results

- Maisonnave, M., Delbianco, F., Tohmé F.A., Maguitman A.G., Milios E. E., 2020, September. Assessing Causality Structures learned from Digital Text Media. In *Proceedings of the 20th ACM symposium on Document engineering*.



Assessing Causality Structures learned from Digital Text Media

Mariano Maisonnave
Depto. de Cs. e Ing. de la Computación, Instituto de Cs. e Ing. de la Computación (ICIC UNS–CONICET)
Bahía Blanca, Argentina
mariano.maisonnave@cs.uns.edu.ar

Fernando Delbianco
Depto. de Economía, Instituto de Matemática de Bahía Blanca (INMABB UNS-CONICET)
Bahía Blanca, Argentina
fernando.delbianco@uns.edu.ar

Fernando Tohmé
Depto. de Economía, Instituto de Matemática de Bahía Blanca (INMABB UNS-CONICET)
Bahía Blanca, Argentina
ftohme@criba.edu.ar

Ana G. Maguitman
Depto. de Cs. e Ing. de la Computación, Instituto de Cs. e Ing. de la Computación (ICIC UNS–CONICET)
Bahía Blanca, Argentina
agm@cs.uns.edu.ar

Evangelos E. Milios
Faculty of Computer Science, Dalhousie University
Halifax, Nova Scotia, Canada
eem@cs.dal.ca

ABSTRACT

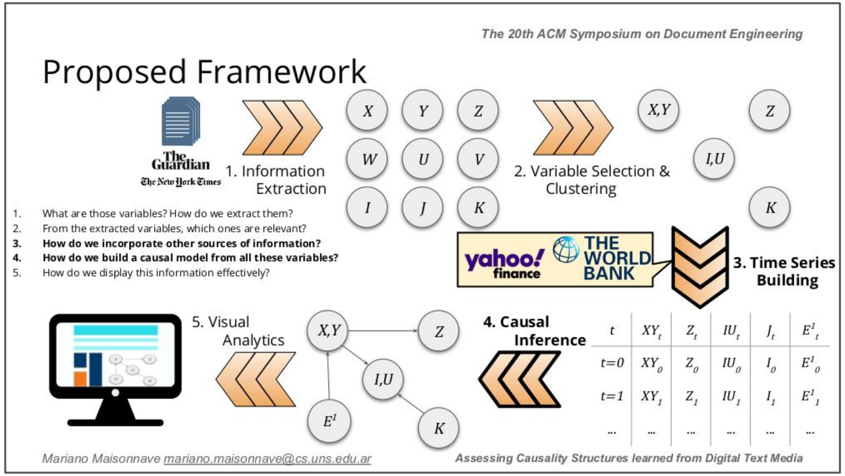
In this paper we describe a framework to uncover potential causal relations between event mentions from streaming text of news media. This framework relies on a dataset of manually labeled events to train a recurrent neural network for event detection. It then creates a time series of event clusters, where clusters are based on BERT contextual word embedding representations of the identified events. Using these time series dataset, we assess four methods based on Granger causality for inferring causal relations. Granger causality is a statistical concept of causality that is based on forecasting. It states that a cause occurs before the effect, and the cause

ACM Reference Format:

Mariano Maisonnave, Fernando Delbianco, Fernando Tohmé, Ana G. Maguitman, and Evangelos E. Milios. 2020. Assessing Causality Structures learned from Digital Text Media. In *ACM Symposium on Document Engineering 2020 (DocEng '20), September 29-October 2, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3395027.3419594

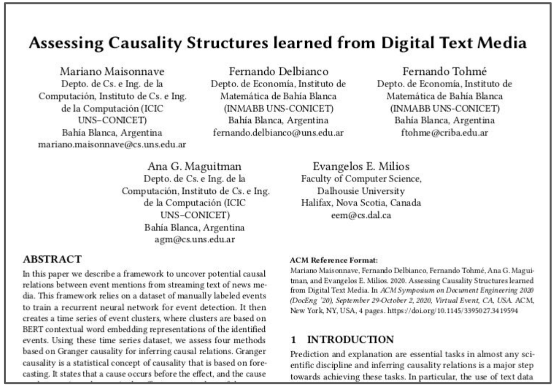
1 INTRODUCTION

Prediction and explanation are essential tasks in almost any scientific discipline and inferring causality relations is a major step towards achieving these tasks. In particular, the use of text data



Results

- Maisonnave, M., Delbianco, F., Tohmé F.A., Maguitman A.G., Milios E. E., 2020, September. Assessing Causality Structures learned from Digital Text Media. In *Proceedings of the 20th ACM symposium on Document engineering*.



Assessing Causality Structures learned from Digital Text Media

Mariano Maisonnave
Depto. de Cs. e Ing. de la Computación, Instituto de Cs. e Ing. de la Computación (ICIC UNS–CONICET)
Bahía Blanca, Argentina
mariano.maisonnave@cs.uns.edu.ar

Fernando Delbianco
Depto. de Economía, Instituto de Matemática de Bahía Blanca (INMABB UNS-CONICET)
Bahía Blanca, Argentina
fernando.delbianco@uns.edu.ar

Fernando Tohmé
Depto. de Economía, Instituto de Matemática de Bahía Blanca (INMABB UNS-CONICET)
Bahía Blanca, Argentina
ftohme@criba.edu.ar

Ana G. Maguitman
Depto. de Cs. e Ing. de la Computación, Instituto de Cs. e Ing. de la Computación (ICIC UNS–CONICET)
Bahía Blanca, Argentina
agm@cs.uns.edu.ar

Evangelos E. Milios
Faculty of Computer Science, Dalhousie University
Halifax, Nova Scotia, Canada
eem@cs.dal.ca

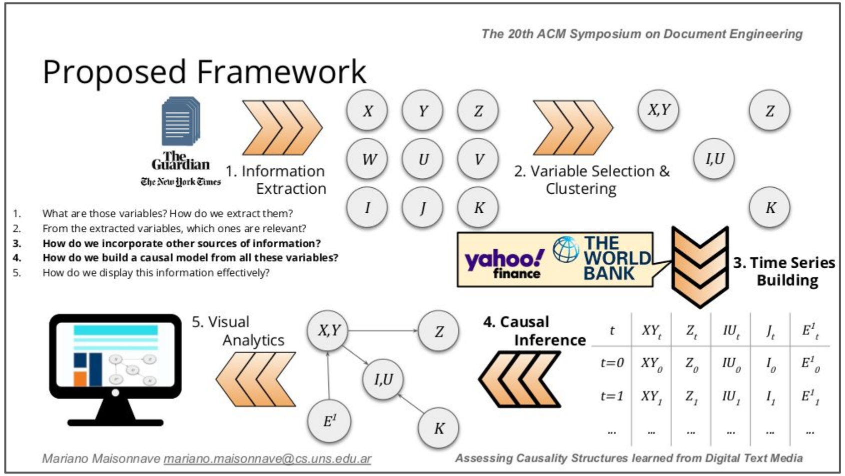
ABSTRACT

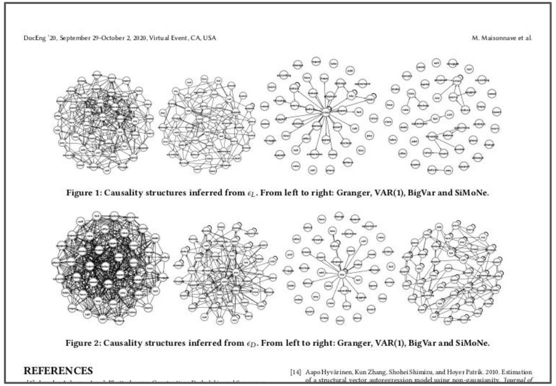
In this paper we describe a framework to uncover potential causal relations between event mentions from streaming text of news media. This framework relies on a dataset of manually labeled events to train a recurrent neural network for event detection. It then creates a time series of event clusters, where clusters are based on BERT contextual word embedding representations of the identified events. Using these time series dataset, we assess four methods based on Granger causality for inferring causal relations. Granger causality is a statistical concept of causality that is based on forecasting. It states that a cause occurs before the effect, and the cause

ACM Reference Format:
Mariano Maisonnave, Fernando Delbianco, Fernando Tohmé, Ana G. Maguitman, and Evangelos E. Milios. 2020. Assessing Causality Structures learned from Digital Text Media. In *ACM Symposium on Document Engineering 2020 (DocEng '20), September 29-October 2, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3395027.3419594

1 INTRODUCTION

Prediction and explanation are essential tasks in almost any scientific discipline and inferring causality relations is a major step towards achieving these tasks. In particular, the use of text data





Thank you!
Muchas gracias!

Feel free to participate in our live Q&A segment and/or contact us as via mail!