JCC2022
JORNADAS CHILENAS DE COMPUTACIÓN

XLI INTERNATIONAL CONFERENCE OF THE CHILEAN COMPUTER SCIENCE SOCIETY, SCCC 2022.

Universidad San Sebastián - Sede Bellavista, Santiago – Chile.

21 al 25 de Noviembre de 2022

Towards automated fact-checking: An exploratory study on the detection of checkable statements in Spanish

Joaquín Saralegui, Antonela Tommasel

JCC2022
JORNADAS CHILENAS DE COMPUTACIÓN

Personal presentation

Joaquín Saralegui

UNICEN, CHEQUEADO

Antonela Tommasel

CONICET - UNICEN

Introduction

"Fake news is **made-up stuff**, masterfully manipulated to **look like credible** journalistic reports that are **easily spread** online to large audiences willing to believe the fictions and spread the word"

Always existed!!

JCC2022
JORNADAS CHILENAS DE COMPUTACIÓN

Introduction

"Fake news is **made-up stuff**, masterfully manipulated to **look like credible** journalistic reports that are **easily spread** online to large audiences willing to believe the fictions and spread the word"

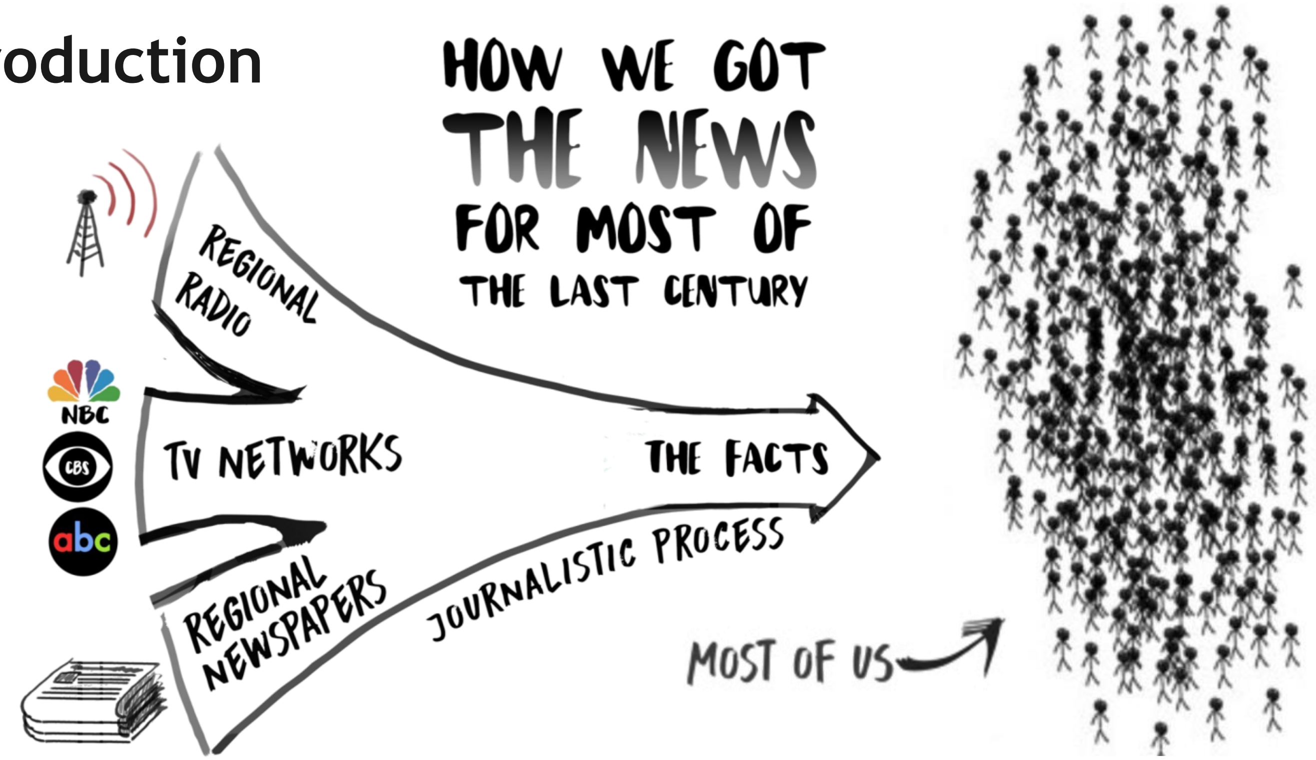
<u>Always existed!!</u>

<u>Social media aggravates the problem!!</u>

A photo or comment that is posted online, and then shared by many people goes viral, spreading from one person to many as quickly as a virus does.

JCC2022
JORNADAS CHILENAS DE COMPUTACIÓN

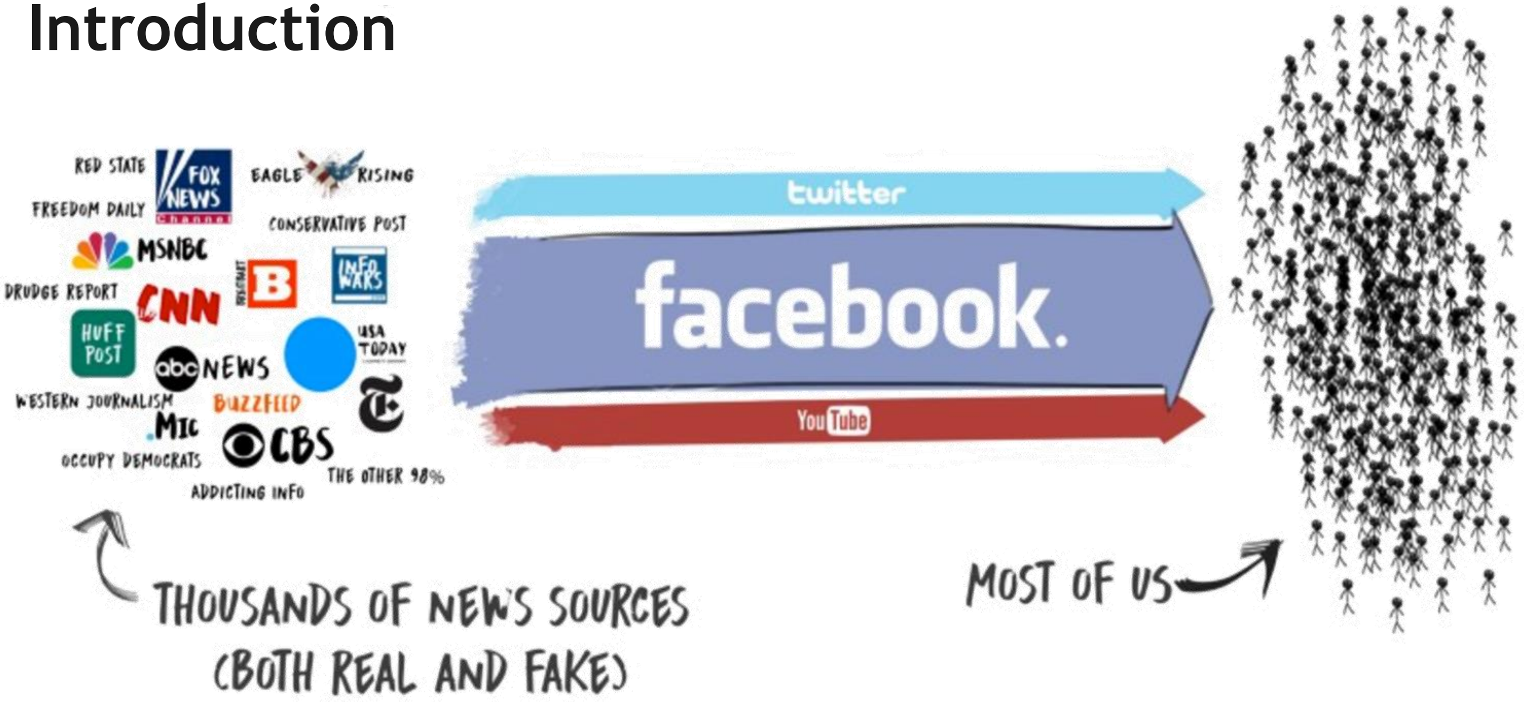
Introduction



https://medium.com/@tobiasrose/empathy-to-democracy-b7f04ab57eee#.100kciuhj

JCC2022
JORNADAS CHILENAS DE COMPUTACIÓN

Introduction



https://medium.com/@tobiasrose/empathy-to-democracy-b7f04ab57eee#.100kciuhj

Introducción

- **Fact-checking organizations** appeared in **1990** in the United States as a response to misinformation spreading.
- **Their goal is to fight misinformation and improve the quality of the public debate** to strenghten the democratic system.
- The **first** fact-checking organization in **Latin America**, **Chequeado**, was founded in **2010**.
- In most cases, they tend to have **limited resources**, while misinformation continues to grow at an increasingly fast pace.
    - The need to **automate as much of the fact-checking process as possible** became an important issue for this organizations.

Motivation - Problem

- **Social media represents the ideal environment for undesirable phenomena!**
  - The dissemination of unwanted or unreliable content, and misinformation.

- **Fake or unreliable content can severely affect society**, posing significant threats to democracies and economy.
  - With the COVID-19 pandemic, health misinformation arose as a threat to public health.

Motivation - Problem

- **Social media represents the ideal environment for undesirable phenomena!**
  - The dissemination of unwanted or unreliable content, and misinformation.

- **Can affect how people perceive content.**
  - Repeated exposure can alter the likelihood of accepting fake content as truth.
  - The line between what is fake or not becomes more uncertain hindering the differentiation between fake and authentic content.
  - The trustworthiness of the entire news ecosystem might be at risk.

Motivation - Problem

It becomes crucial to fact-check the factuality and authenticity of the shared information.

- In addition to claim verification, one of the key tasks in such verification process is to **determine which statements can be fact-checked.**

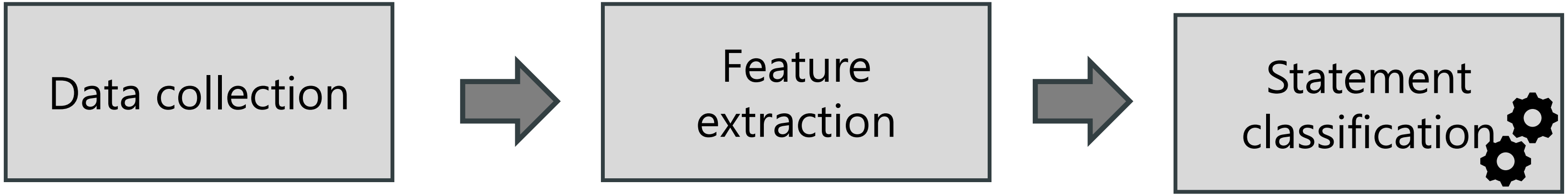
Motivation - Problem

It becomes crucial to fact-check the factuality and authenticity of the shared information.

- In addition to claim verification, one of the key tasks in such verification process is to **determine which statements can be fact-checked.**

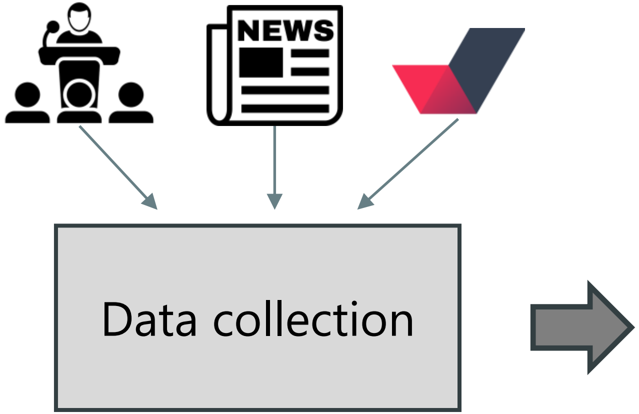
We evaluate the performance of different approaches for identifying checkable statements in the Spanish language.

Proposal


Data collection
Feature extraction
Statement classification

JCC2022
JORNADAS CHILENAS DE COMPUTACIÓN

Proposal - Data collection



"In 2022, inflation has decreased in Argentina."
"I think X will happen."
"I promise inflation will decrease."

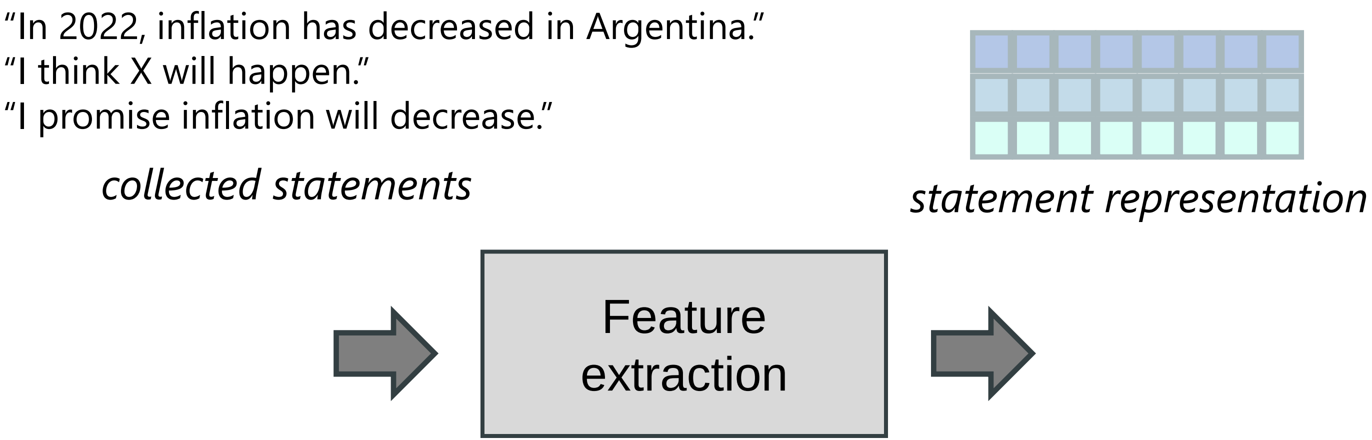
collected statements

- **Three** data sources were considered:
  - **Presidential speeches** are representative of political speeches (2300).
  - **Fact-checks** refer to statements that have already been factchecked (1300).
  - **News** covering diverse contexts (1400).

- Labelling was performed following Chequeado guidelines, involving the work of a professional fact-checker.

- We obtained **4958** statements.
  - **39% fact-checkable.**
  - **61% non-checkable.**

Proposal - Feature extraction



- The extracted features can be divided into two groups:
  - **Traditional representation** (3500 features approx., we selected half based on χ^2).
  - **Semantic representations** (512 features).

Proposal - Feature extraction: Traditional representation

- Mainly based on the **lexical analysis of texts**.
- Texts are usually represented by the **Bag-of-Words (BoW)** model.
  - Each **word** or term is **represented as an independent feature**.
- This representation **disregards all grammar considerations**, context, and word order but keeps the information about the frequency of each term.
- A **binary weighting scheme** was considered.

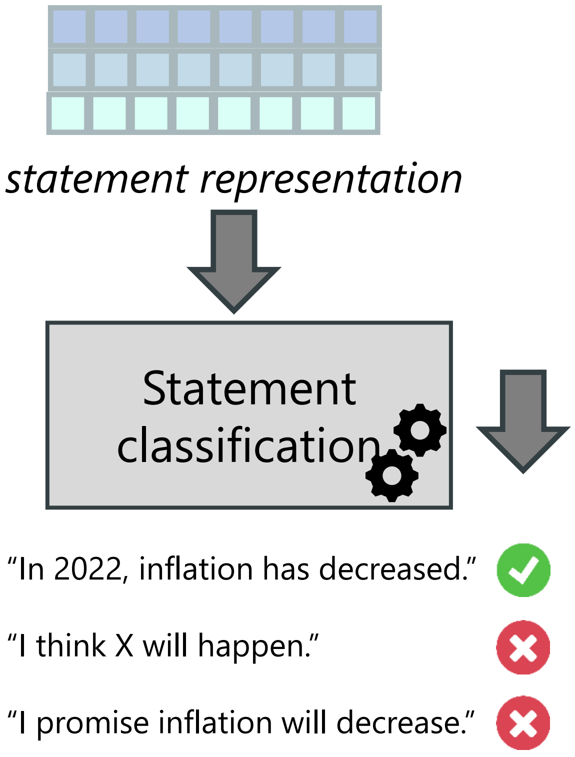
Lemmas	Named entities	POS tags

Proposal - Feature extraction: Semantic representation

- The **BoW model** has a few drawbacks.
  - It can create **large feature spaces**, which could become sparse.
  - It **assumes that words are independent of each other**, ignoring potential semantic relationships between them.
  - It **requires a large number of instances** to extract relevant information.
- **"word embeddings"** can convert texts into numeric vectors aiming at **capturing** words' **semantics**.
- Unlike word embeddings, **sentence embeddings** better capture polysemy, word order, out-of-vocabulary words, and sentence forms.
- We used the multilingual **Universal Sentence Encoder (USE)** to generate a 512 fixed dimension and dense representation of each statement.

Proposal - Statement classification

- We define the detection of checkable statements as a **binary classification task** in which statements are classified as “checkable” or “non-checkable”.
- We chose the **most commonly used classification techniques**:
  - Multinomial Naïve Bayes (NMB).
  - Random Forest (RF).
  - Support Vector Machines (SVM).
  - Logistic Regression (LR).
- The collected data was divided into train and test sets following a **k-fold stratified cross-validation model**, setting k to 5.
  - There was no need to account for temporal data relation.



Proposal - Implementation details

- All processing was implemented in **Python**.
- **SpaCy** was selected for NLP processing as it provided a more optimized pipeline allowing for faster processing and a good coverage of the Spanish language.
- Classifiers were implemented using **Sklearn**.
- The **same data partitions** were used for all evaluations.
- Performance was evaluated considering the macro and per class **Precision, Recall and F-Measure**.

Results

		Precision		Recall		F-Measure	
		Macro	Check. class	Macro	Check. class	Macro	Check. class
Random		0.49 ±0.02	0.39 ±0.02	0.50 ±0.02	0.5 ±0.04	0.49 ±0.02	0.43 ±0.03
Traditional representation	LR	0.83 ±0.01	0.79 ±0.02	0.83 ±0.01	0.78 ±0.02	0.83 ±0.01	<u>0.79</u> ±0.02
	MNB	0.83 ±0.02	0.79 ±0.03	0.82 ±0.01	0.78 ±0.01	0.83 ±0.01	0.78 ±0.02
	RF	0.84 ±0.01	0.84 ±0.02	0.81 ±0.01	0.69 ±0.02	0.82 ±0.01	0.76 ±0.01
	SVM	0.85 ±0.01	0.85 ±0.02	0.82 ±0.01	0.72 ±0.01	0.83 ±0.01	0.78 ±0.01
Semantic representation	LR	0.83 ±0.01	0.78 ±0.01	<u>0.84</u> ±0.02	**0.83** ±0.03	<u>0.84</u> ±0.01	**0.80** ±0.01
	MNB	0.84 ±0.01	**0.90** ±0.03	0.77 ±0.01	0.59 ±0.02	0.79 ±0.01	0.71 ±0.02
	RF	0.84 ±0.01	0.87 ±0.02	0.80 ±0.01	0.66 ±0.02	0.81 ±0.01	0.75 ±0.01
	SVM	<u>0.86</u> ±0.01	0.86 ±0.02	**0.85** ±0.01	0.77 ±0.02	**0.85** ±0.01	0.82 ±0.01
Combined representation	LR	0.84 ±0.01	0.80 ±0.01	<u>0.84</u> ±0.01	<u>0.80</u> ±0.01	<u>0.84</u> ±0.01	**0.80** ±0.01
	MNB	0.84 ±0.01	0.82 ±0.02	0.83 ±0.01	0.79 ±0.02	<u>0.84</u> ±0.01	<u>0.80</u> ±0.01
	RF	<u>0.86</u> ±0.02	**0.90** ±0.03	0.80 ±0.01	0.65 ±0.01	0.82 ±0.02	0.75 ±0.03
	SVM	**0.87** ±0.01	<u>0.88</u> ±0.03	<u>0.84</u> ±0.01	074 ±0.02	**0.85** ±0.01	**0.80** ±0.02

JCC2022 JORNADAS CHILENAS DE COMPUTACIÓN

Results

		Precision		Recall		F-Measure	
		Macro	Check. class	Macro	Check. class	Macro	Check. class
Random		0.49 ±0.02	0.39 ±0.02	0.50 ±0.02	0.5 ±0.04	0.49 ±0.02	0.43 ±0.03
Traditional representation	LR	0.83 ±0.01	0.79 ±0.02	0.83 ±0.01	0.78 ±0.02	0.83 ±0.01	0.79 ±0.02
	MNB	0.83 ±0.02	0.79 ±0.03	0.82 ±0.01	0.78 ±0.01	0.83 ±0.01	0.78 ±0.02
	RF	0.84 ±0.01	0.84 ±0.02	0.81 ±0.01	0.69 ±0.02	0.82 ±0.01	0.76 ±0.01
	SVM	0.85 ±0.01	0.85 ±0.02	0.82 ±0.01	0.72 ±0.01	0.83 ±0.01	0.78 ±0.01

- LR and MNB achieved similar results for the three metrics, while **RF and SVM achieved a slightly higher precision than recall**.
- Macro metrics were higher than those for the checkable counterpart.
  - Differences were more noticeable for **RF and SVM**, which achieved both the **highest precision and lowest recall**.
- **Classifiers are more confident that the identified checkable statements are actually checkable, at the expense of identifying fewer of them.**

JCC2022
JORNADAS CHILENAS DE COMPUTACIÓN

Results

		Precision		Recall		F-Measure	
		Macro	Check. class	Macro	Check. class	Macro	Check. class
Semantic representation	LR	0.83 ±0.01	0.78 ±0.01	0.84 ±0.02	**0.83** ±0.03	0.84 ±0.01	**0.80** ±0.01
	MNB	0.84 ±0.01	**0.90** ±0.03	0.77 ±0.01	0.59 ±0.02	0.79 ±0.01	0.71 ±0.02
	RF	0.84 ±0.01	0.87 ±0.02	0.80 ±0.01	0.66 ±0.02	0.81 ±0.01	0.75 ±0.01
	SVM	0.86 ±0.01	0.86 ±0.02	**0.85** ±0.01	0.77 ±0.02	**0.85** ±0.01	0.82 ±0.01

- Macro metrics: LR, RF and SVM achieved similar results than for the traditional representation, while MNB decreased its recall by 7%.

- Checkable class: recall decreased 18% on average for MNB and RF, while slightly increased 6% on average for LR and SVM.

- **LR and SVM increased recall while maintaining precision.**

Results

- For LR, differences in the checkable metrics with the semantic representations were not statistically significant.
- Despite its improvements, **MNB still had slightly worse performance than LR and SVM** in terms of F-measure, precision (SVM) and recall (LR).
- **RF kept the tendency for high precision but low recall, making the model unsuitable for the task.**
- For SVM, precision differences were statistically significant, while recall differences were not.

		Precision		Recall		F-Measure	
		Macro	Check. class	Macro	Check. class	Macro	Check. class
Combined representation	LR	0.84 ±0.01	0.80 ±0.01	0.84 ±0.01	0.80 ±0.01	0.84 ±0.01	**0.80** ±0.01
	MNB	0.84 ±0.01	0.82 ±0.02	0.83 ±0.01	0.79 ±0.02	0.84 ±0.01	0.80 ±0.01
	RF	0.86 ±0.02	**0.90** ±0.03	0.80 ±0.01	0.65 ±0.01	0.82 ±0.02	0.75 ±0.03
	SVM	**0.87** ±0.01	0.88 ±0.03	0.84 ±0.01	074 ±0.02	**0.85** ±0.01	**0.80** ±0.02

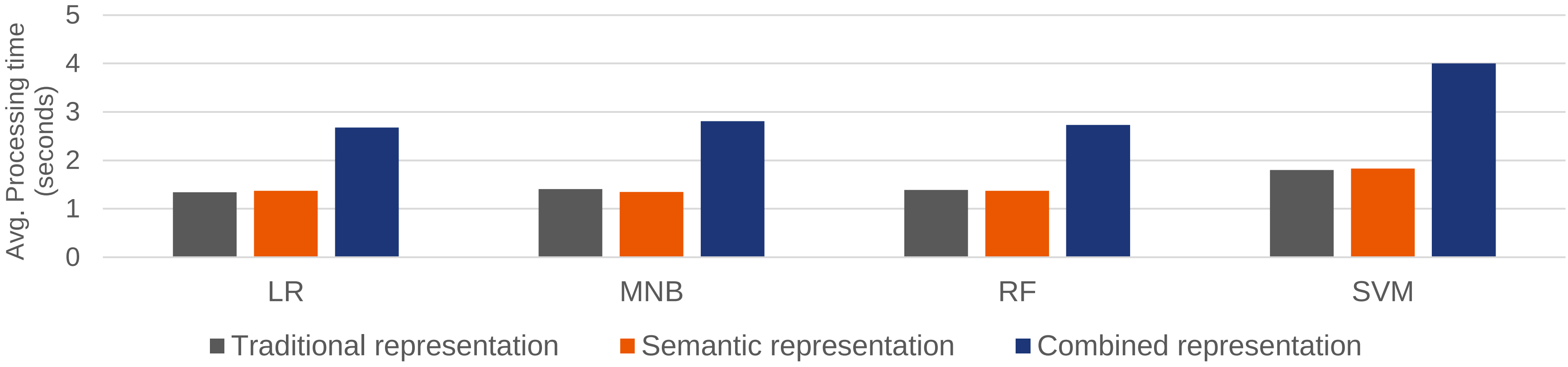
JCC2022
JORNADAS CHILENAS DE COMPUTACIÓN

Results

		Precision		Recall		F-Measure	
		Macro	Check. class	Macro	Check. class	Macro	Check. class
Random		0.49 ±0.02	0.39 ±0.02	0.50 ±0.02	0.5 ±0.04	0.49 ±0.02	0.43 ±0.03
Traditional representation	LR	0.83 ±0.01	0.79 ±0.02	0.83 ±0.01	0.78 ±0.02	0.83 ±0.01	0.79 ±0.02
	MNB	0.83 ±0.02	0.79 ±0.03	0.82 ±0.01	0.78 ±0.01	0.83 ±0.01	0.78 ±0.02

- The evaluated representations and models **obtained similar results** to both the **human study** and the **state-of-the-art models for the English language**.
- **LR and SVM were the best performing models** for the three evaluated representations.
- **LR** would be able to **identify more relevant statements**.
- **SVM** would be able to **reduce the number of incorrectly identified statements** (reducing the load of human checkers) while missing some relevant statements.

Results - Time comparison



- **Combining the features doubled the required time.**
- SVM was approximately 33% slower than the other models.
- On average, **models can process approximately 500 statements per second**.
  - Models **allow for the real-time processing** of statements.

Conclusions

- This study tackled the **identification of checkable statements in Spanish** by evaluating **different combinations of features and classifiers**.
- The models achieved comparable results to state-of-the-art techniques for English text.
- The best performing **model is publicly available** and, at the time of writing, used by different fact-checking organizations in Latin America.

- Additional evaluations could be performed over data collections from different Spanish variations.
- A more in-depth study of the contribution of each feature type to model performance is needed.
- Extend the search of checkable statements to other domains, such as messaging platforms or even video transcripts.

JCC2022
JORNADAS CHILENAS DE COMPUTACIÓN

Thanks!
Questions?

antonela.tommasel@isistan.unicen.edu.ar

JCC2022
JORNADAS CHILENAS DE COMPUTACIÓN

XLI INTERNATIONAL CONFERENCE OF THE CHILEAN COMPUTER SCIENCE SOCIETY, SCCC 2022.

Universidad San Sebastián - Sede Bellavista, Santiago – Chile.

21 al 25 de Noviembre de 2022

Towards automated fact-checking: An exploratory study on the detection of checkable statements in Spanish

Joaquín Saralegui, Antonela Tommasel