Towards an automated fact checking process

...

chequeado

LA VERIFICACIÓN
DEL DISCURSO PÚBLICO

About me

- Mariano Falcón
- Software Engineer
- Data visualisations, NLP
- @falconius

About Chequeado

- Non partisan, non for profit organization from Argentina
- First fact checking site in Latin America
- Founded in 2009, online since 2010

What’s fact checking?

It’s the process of taking a claim from the public domain and asserting their trueness.

Why automating it?

- Expanding our reach.

We are not alone

- FullFact (in partnership).
- Politifact (Claimbuster).

Chequeado's fact checking steps (summarized)

1- Selection of the statement to check.

2- Consult sources.

3- Contextualize information.

4- Ranking.

Chequeado’s fact checking steps (summarized)

1- Selection of the statement to check **-> 1) Data collection 2) Fact Check detection**

2- Consult sources.

3- Contextualize information.

4- Ranking.

Data collection

Text based

- Online newspapers
- Social media
- Transcripts (congress, debates, etc)

Audiovisual based

- TV
- Radio
- Youtube/Video streaming services

Data collection

Text based

Time consuming, but easy to solve

- Website scraping
- Feed parsing
- Pdf parsing
- 3rd party API's consumption

Audiovisual based

Difficult to get the data without significant noise

- Closed caption extraction
- Speech recognition software or services
- Clipping agencies

Help needed here!

What makes a claim “fact checkable”?

- Contain comparable data:

  “Inflation has raised ten percent in the last year.”

- Not opinions nor expressing wishes:

  “We will invest 3MM in the next two years.”

Do you already know about NLP?

- Download the exercise code (check the README for instructions): https://github.com/chequeado/autofact

What is NLP?

Natural Language Processing

Use computers to analyze natural language.

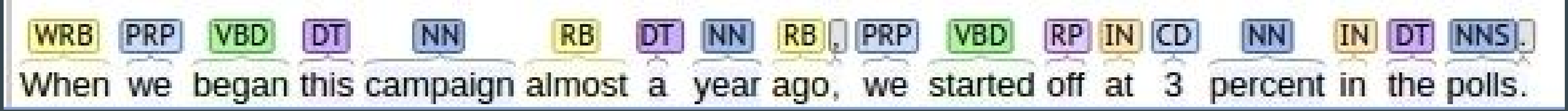
NLP - Lemmatization

Reduce words to their normal form.

- Being, Been -> **be**
- increased, increasing, increase -> **increase**
- Rose, rise, risen, rising -> **rise**

NLP - PoS (Part of Speech) Tagging

Classifying words into their parts of speech.



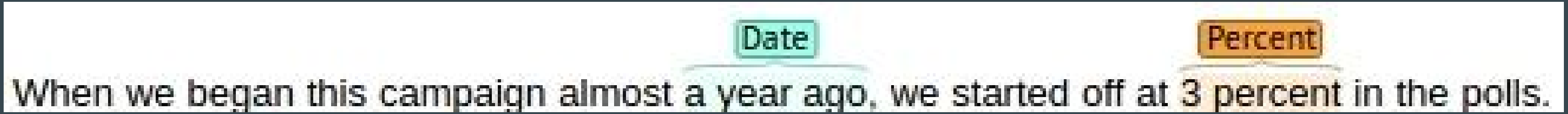
WRB	PRP	VBD	DT	NN	RB	DT	NN	RB	,	PRP	VBD	RP	IN	CD	NN	IN	DT	NNS	.
When	we	began	this	campaign	almost	a	year	ago	,	we	started	off	at	3	percent	in	the	polls	.

Some tag descriptions:

NN: Noun, singular or mass
RB: Adverb
VBD: Verb, past tense

NLP - NER (Named Entity Recognition)

Label sequences of words into pre-defined categories.



When we began this campaign almost a year ago, we started off at 3 percent in the polls.

Entity types:

- DATE
- PERCENT
- MONEY
- PERSON
- LOCATION
- ORGANIZATION
- DURATION
- NUMBER
- ORDINAL
- MISC

Having a classified dataset of claims

- Hillary Clinton : When I was secretary of state, we actually increased American exports globally 30 percent -> **Fact checkable**
- Donald Trump : We're a country that owes $20 trillion. -> **Fact checkable**
- Donald Trump : Under my plan, I'll be reducing taxes tremendously, from 35 percent to 15 percent for companies, small and big businesses. -> **Not fact checkable**
- Hillary Clinton : Going from $7.25 to $12 is a huge difference. -> **Not fact checkable**

Two approaches to detect fact checkable sentences

We have none or a very small set of classified data

Create custom rules from experience and observation:

- Verbal tenses
- Numerals
- Dates
- Specific words

Yes, we have some already classified data

Use the tagged data to train a predictive model.

We have none or a very small set of classified data

Create custom rules from experience and observation:

- Verbal tenses
- Numerals
- Dates
- Specific words

Let’s try it!

Yes, we have some already classified data

Use the tagged data to train a predictive model.

- Repository url: https://github.com/chequeado/autofact
- More info on the README