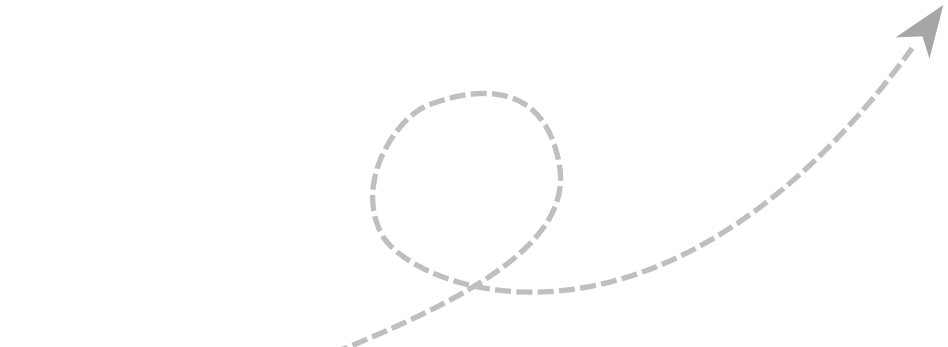
# CO-TRAINING IMPLEMENTATION FOR OBJECTIVITY DETECTION

## Domain – Sport News

Cuevas Saavedra, Vladimir Enrique

Segura Tinoco, German Andrés



# Application

Why this application is useful and what is going to resolve?


## Main Applications

For editors in chief it could be very useful to have a tool that helps them know which sports news are objective and which are subjective.

For people that read digital news, would be helpful to have a way to know if some sport articles was written objectively or not.

# Motivation

Why we picked this idea and implementation?

## First Idea

Everyday 78% of the young people from Latin America read sport news, 51% of those read before starting any other activities. They want reliable and impartial information.

## Second Idea

It's interesting to explore the subjectivity analysis from sport news and to avoid the yellow press

## Third Idea

Sports news is an industry that moves millions of dollars every day. We think that this idea can be very profitable










## Why Sports News?

There is a big euphoria about sports, sports basically is anything that humans find entertaining or any activity for engaged in for relaxation and amusement. It can come from Chess and Video Games to Mixed Martial Arts and Formula 1; and no matter what you think you have a sport that you'd like and from time to time we would like to read about the competition that can have great impact on the reader, then is when is required the impartiality of the computers to guarantee that we are reading something that is providing facts.

# Approaches to be Used

Which approach, technique and algorithms can help us to resolve the problem?

## Associated Tasks

- Classification
- Subjectivity Analysis
- Information Retrieval

## Algorithms

- Naïve Bayes and maybe SVM
- Co-training with Late Fusion

## Approach

- Implementation project
- Semi-Supervised Learning for Classification
- Use 2 views: Raw data view and structured-labeled data view

# Dataset Description

What are going to be our inputs and data?

**Dataset is composed of two parts:**

- **RAW Data:** It contains several text files, each of them with an article from a real magazine, news, or similar. It is plain text and will be processed to extract the TF-IDF and generate new features.
- **Structured data:** It is a file that contains an analysis for each of the files in raw data, it contains about 58 features and a conclusion on the article stating if it is objective or subjective. Some of the features are retrieved using the Stanford POS tagger and the tags are as defined in Penn Treebank Project which basically are part-of-speech.

**Source**

- UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)

# Related Work

**List of previous research**

- **Research 1:** Nadine Hajj, Yara Rizk, and Mariette Awad, 'A Subjectivity Classification Framework for Sports Articles using Cortical Algorithms for Feature Selection,' Springer Neural Computing and Applications, 2018.
    - Research that created a framework for classification of sports articles, achieving an 85.6% of accuracy on a 4-fold cross validation 40% reduction in features using CA*

- **Research 2:** Yara Rizk, and Mariette Awad, 'Syntactic Genetic Algorithm for a Subjectivity Analysis of Sports Articles,' International Conference on Cybernetic Intelligent Systems, Limerick, Ireland, 2012.
    - Research to apply artificial intelligence to automate classification of sport articles, they proposed the use of genetic algorithms with syntactic features for subjective content analysis

# Thanks