European
Commission

Web Intelligence for Measuring Emerging
Economic Trends: the Drone Industry

Searching the Web for the Drone Industry: Classifying Websites in Multiple Countries and Languages with a Single Model

SDSS 2022 conference, Pittsburgh

Session CS14

9 June 2022

Piet Daas*, Statistics Netherlands & Eindhoven Univ. of Technology*
Blanca de Miguel, Universitat Politecnica de Valencia
Maria de Miguel, Universitat Politecnica de Valencia

Web Intelligence for Drones: Objectives and starting point

- Is it possible to collect information from the web on businesses based in Europe that have their main activity in the civil Drone sector?

- **Starting points of the work:**
  - Previous research on web scraping of businesses information in the context of official statistics (a number of EU-projects)
  - Exploratory sector analyses (country specific: Spain, Ireland and Italy)
    - Manually collected and verified lists of drone domain names for Spain (1097) and Italy (686)
  - Testing of different web search strategies
  - ***Develop a model to preselect and/or identify Drone companies in large datasets***

Search engine based approach

- **Web search strategy** to identify the universe/population of drone businesses based on the world wide web for a country

- Search features/criteria:

  - Search queries: search words + composition of queries (many!!!!)

  - 2 languages for each country: national language + English (for Spain, Ireland, Italy)

  - 6 search engines: Google, Bing, DuckDuckGo, Yahoo, AOL, Ask

Web search strategy: results

Examples of search queries

1. drone company spain & drones empresa espana
2. (drone OR rpas OR uav OR uas) registration spain & dron OR rpas OR uav OR uas) registro espana

- Search for individual websites
- Search for websites containing overviews of Drone companies
- Search for (PDF-)files containing URLs of Drone companies
- Search for (PDF-)files containing names of Drone companies

- This resulted in *many* URLs that could refer to websites of Drone companies

Web Intelligence for Measuring Emerging Economic Trends: the Drone Industry

Results of web search strategy

Script	Links found	Spain EN/Spanish	Ireland EN	Italy EN/Italian
Step 1	Web-links	33 274 / 31 546	29 958	29 058 / 21 848
	PDF-links	1027 / 24	878	768 / 485
Step 2	Web-links (a)	22 608 / 6542	56 513	53 937 / 56 906
	Web-links (b)	134 / 306	182	105 / 356
	PDF-links	1861 / 9541	1974	2421 / 12 281
Step 3	Web-links	5886 / 1957	6115	7011 / 3996
	Name-based	7065	2816	34 185
Step 4a	Web-links	47 980 / 49 201	48 950	115 107 / 115 253
Step 4b	Web-links	46 981 / 47 101	14 568	112 901 / 112 066
Total uniqueURLs	Combined	**26 067**	**14 568**	**53 781**

Which of these URLs are of a Drone company?

European Commission

Develop a Classification model

- **Starting points:**
  - A dataset is available containing 1.097 Spanish drone websites
  - The URLs collected for Spain will certainly contain non-Drone Websites

- **Determine which preprocessing steps and classification algorithms produce the most promising results**
  - A supervised ML-task
  - Tried both positive and unknown (PUlearn) based and a whole range of positive and negative based approaches

Develop a Classification model (2)

1. PUlearning based approach (positive and unknown cases)

- Spanish drone list as positive examples (1.097)
- Results of search approach for Spain as unknown input (sample from ~26.000)
- Looked OK at first, No stemming and translating all Spanish words to English improved accuracy (on positive cases). Max. accuracy of 87%
- However, applying model to an unseen part of the unknown dataset resulted in:
  - Probabilities > 1 (max 1.7) ?
  - Manual inspection showed that Negative classified cases contained obvious Drone websites and Positive classified cases contained obvious non-Drone websites

2. Traditional ML-classification approach (based on positive AND negative cases)

- Manually checked 3000 randomly selected URLs from search approach to obtain non-Drone websites
- Determined effect of various preprocessing steps, including translating Spanish to English words, and compared a whole range of ML-classifiers (included in scikit-learn).

Develop a Classification model (3)

2a. Preprocessing

- Language detection (ES,EN), stopwords removal, remove numbers, remove punctuation marks
- Stemming (?), Remove words of 2 or 3 character lengths (?)
- Effect of translating Spanish to English words (?)

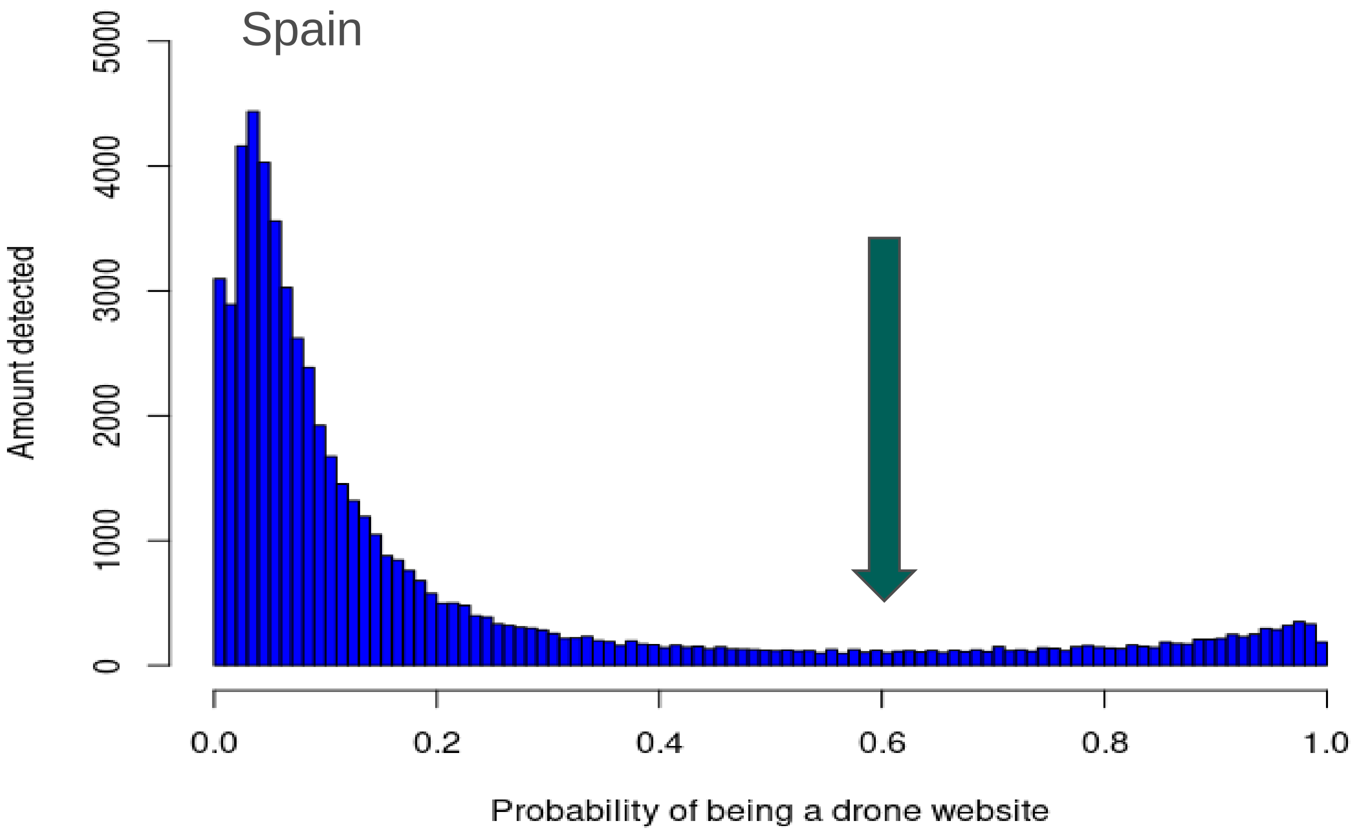
Because many texts (websites) needed to be translated, ‘Apertium’ was used

- Open source/free, off-line translation
- Spanish to English translation is OK
- Indicates which words are not well translated (if needed, added the correct translations)

The Classification Model (4)

- Model properties
  - Best algorithm *Log. Reg.* L2-norm (Acc. 87%, Prec. 76%, Recall 93%)
  - Best preprocessing choices: Mindf 100, Maxdf 2000, Min. char 3, No stemming, Translating Spa -> Eng,
  - Model contains 1568 Features (of which 1559 are words, others are inclusion of drone synonyms in URL and on webpage)
  - Only English words are included as features
  - Model is especially good in identifying non-Drone websites (Acc. 93%)
  - Produces either a 0/1 or the probability of being a drone website (value between 0-1)
  - Manual checking of independent sample (100), in various probability ranges, by Spanish Drone experts revealed an Acc. of 85%
    - Best cut-off value is 0.6

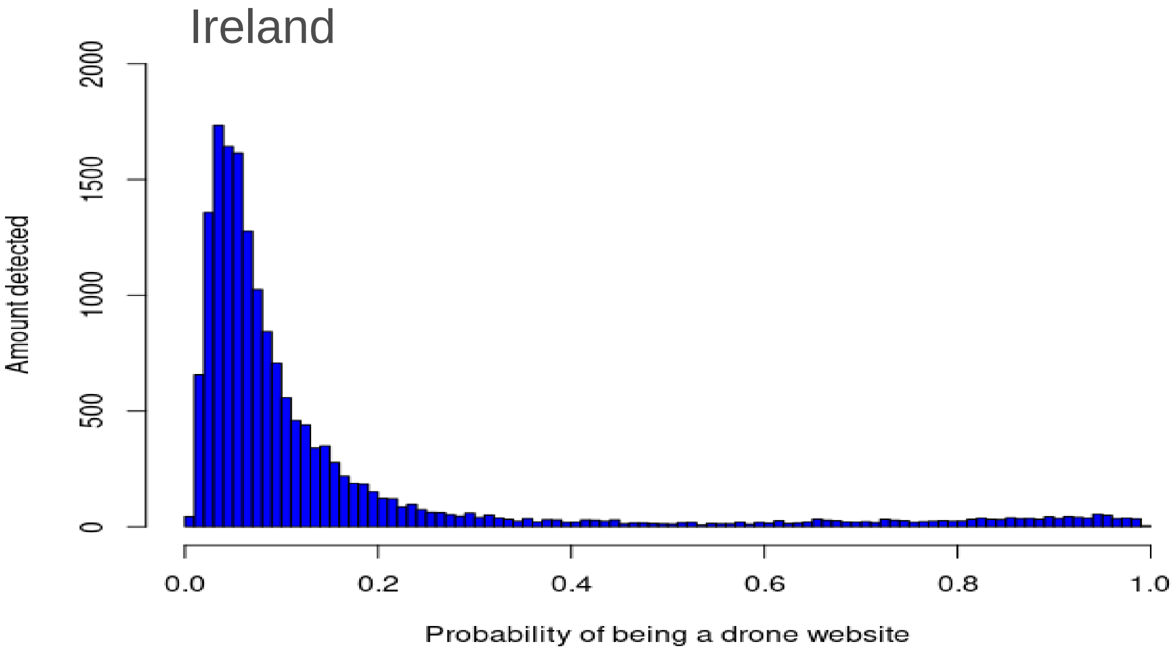
Applying model on all Spanish URLs



In the end a total of 461 Drone websites were found

Applying model to Irish dataset

- Model results
  - No language issues, all websites were written in English (no Gaelic drone websites included)
  - Classification results were manually inspected by experts (random samples on various ranges)
    - Acc. 86%, Prec. 72%, Recall 100%, best cut-off value 0.6



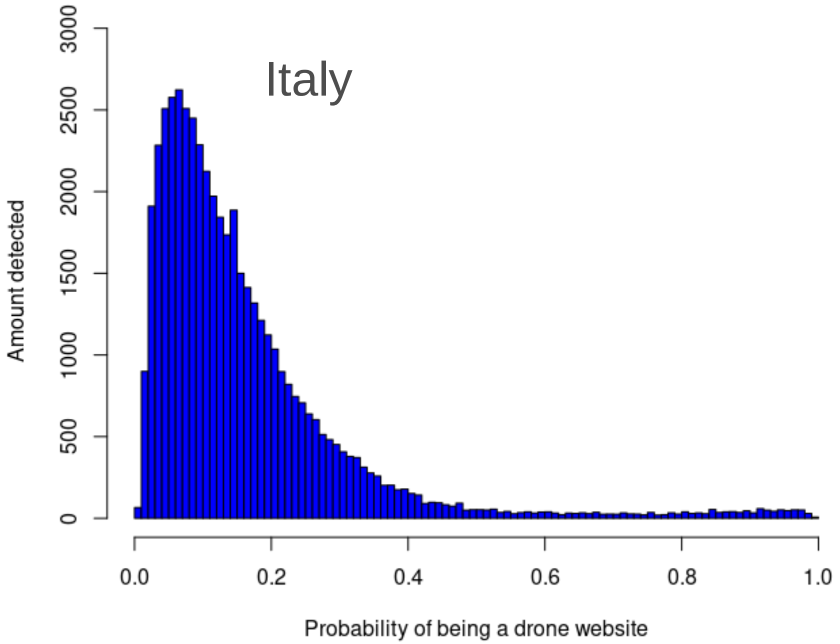
In the end a total of 66 Drone websites were identified

The Classification Model: Italy

- How to apply model to non-English websites?
  - Model includes ONLY English words
  - **Step1:** Create translation word list (Eng -> Ita)
  - Challenging, because some important words in Spanish dataset were written in English (so not all English words in model *had to be* translated!)
    - ‘Web’, ‘cookie’, ‘cookies’, ‘log’ are NOT translated
    - Deal with male/female versions of words (one -> una, uno)
    - ‘fly’ and ‘unmanned’ -> ‘volare’ and ‘senza pilota’ (NOT ‘mosca’ and ‘senza equipaggio’)
    - Any additional adjustments decreased results!
  - **Step 2**: Prior to applying the model, all words included in translation file that occurred on Italian webpages needed to be translated to English
  - **Step 3:** Create features added for country (lang_feat, drone word specific features)

Applying model to Italian dataset

- Model results
  - Test the Ita-Eng translation of the models' words on the 686 identified Italian drone websites (Acc. 85%)
  - Applied to all Italian websites found, followed by manually inspected by experts (random samples in various ranges)
    - Acc. 82%, Prec. 67%, Recall 97%, best cut-off value 0.6



In the end a total of 353 Drone websites were identified

Conclusions (for model)

- Model trained on Spanish drone websites – provided valid results for Ireland and Italy (accuracies of 82-86%). Recall is high.
- Model is particularly well suited to remove non-Drone websites from large numbers of URLs (high accuracy on negative cases and high recall on positive cases)
- The model could be applied to websites in other countries when:

  1) websites are written in English or when a ‘correct’ translation list has been created

  2) the features used on drone websites for the country studied are comparable to those used in Spain, Ireland and/or Italy

Thank you

Dissemination of project's results

- Full project's results are available on CROS (Collaboration and Research for Official Statistics): Web Intelligence for Drones
- Scripts will be published on Eurostat GitHub WIH Drones

European Commission