Journal of Data Science 0 (0), 1–14
??? 2023

DOI: 10.6339/23-JDS1087
Data Science in Action

Identifying Drone Web Sites in Multiple Countries and Languages with a Single Model

Piet Daas[1,2,*,†], Blanca de Miguel[3], and Maria de Miguel[3]

[1]*De Groene Loper 5, 5612AZ Eindhoven, Eindhoven University of Technology, the Netherlands*
[2]*CBS-weg 11, 6412EX, Heerlen, Statistics Netherlands, the Netherlands*
[3]*Camí de Vera, s/n 46022 Valencia, Universitat Politècnica de València, Spain*

Abstract

A text-based, bag-of-words, model was developed to identify drone company websites for multiple European countries in different languages. A collection of Spanish drone and non-drone websites was used for initial model development. Various classification methods were compared. Supervised logistic regression (L2-norm) performed best with an accuracy of 87% on the unseen test set. The accuracy of the later model improved to 88% when it was trained on texts in which all Spanish words were translated into English. Retraining the model on texts in which all typical Spanish words, such as names of cities and regions, and words indicative for specific periods in time, such as the months of the year and days of the week, were removed did not affect the overall performance of the model and made it more generally applicable. Applying the cleaned, completely English word-based, model to a collection of Irish and Italian drone and non-drone websites revealed, after manual inspection, that it was able to detect drone websites in those countries with an accuracy of 82 and 86%, respectively. The classification of Italian texts required the creation of a translation list in which all 1560 English word-based features in the model were translated to their Italian analogs. Because the model had a very high recall, 93, 100, and 97% on Spanish, Irish and Italian drone websites respectively, it was particularly well suited to select potential drone websites in large collections of websites.

Keywords *bag of words; classification model; multiple languages; text*

1 Introduction

Nowadays an estimated number of at least 5.13 billion web pages are available online (De Kunder, 2022). A considerable part of those pages is produced by companies. These pages contain texts that may enable very interesting applications, such as providing novel insights on the activities of companies (Gökk et al., 2015; Aweisi et al., 2021), to inform policymakers (Höchtl et al., 2015), in official statistics production (Florescu et al., 2014), and to obtain insights on new and emerging economic trends (GOPA, 2021b). However, extracting relevant and reliable information from website texts in a reproducible way is not an easy task (Antonacopoulos and Hu, 2003; Kitchin, 2015).

Website texts have been investigated by many researchers and examples can be found in 'Web mining' books, such as Larose and Markov (2007), and in more text-oriented Web mining

[*] Corresponding author. Email: p.j.h.daas@tue.nl.

[†] The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Received July 25, 2022; Accepted January 17, 2023

books, such as Song and Wu (2008). To obtain more information on the activities of companies, several applications have been described. Daas and van der Doef (2020) found that website texts can be used to identify innovative companies. In Daas and de Wolf (2021) the application of website texts to identify platform economy companies and companies active in the area of Artificial Intelligence are mentioned. Kühnemann et al. (2020) studied the use of website texts to determine the main economic activity of a company. In addition, websites have been used to collect product prices (Powell et al., 2018), obtain online job advertisements (Beręsewicz and Pater, 2021), update information on enterprises, or produce new output (ESSnet, 2020).

In this paper, we focus on the detection of companies active in new and emerging economic areas. This is traditionally done by consulting experts and often includes collecting data via a survey. The downsides of such an approach are that it is slow, that it usually takes a few iterations to obtain a complete overview, and that it puts an administrative burden on companies as they have to fill in a questionnaire. Here, we propose an alternative approach, one that focuses on using information included in website texts. With the drone industry as an example, website texts were used to develop a model capable of determining if a company is active in this new and emerging economic area. Here, being active in the drone industry means that companies buy from suppliers of drone goods and services and/or offer drone products and services to their buyers (based on Rothaermel (2019)).

We also looked at developing a model capable of classifying web pages written in different languages. This is highly relevant when one wants to study a new and emerging economic area in multiple countries. In the Natural Language Processing literature, the most common approaches used to create text-based classification models are by using i) Word count-based features, ii) Word embeddings, or iii) Transformers (Kowsari et al., 2019). In each case, the model developed requires numerical input, which means that – in some way or another – texts have to be converted into numbers (Gentzkow et al., 2019). For Word count-based features this is done by splitting texts into words, or multi-sets of n-gram words, to which weights, such as counts or frequencies, are assigned. Since words differ per language, a multilingual variant of such an approach needs to capture the relation between words with the same meaning in each language. Word embedding is a technique in which each word or phrase is mapped to a N dimensional vector of real numbers. It has the advantage that words of similar meaning are expressed as similar numerical vectors. This makes this approach potentially less language dependent (Almeida and Xexéo, 2019). Transformers extend the Word embeddings approach by additionally incorporating the context in which words are used. Usually, pre-trained neural networks are used of which the Bidirectional Encoder Representations from Transformers (BERT) developed by Google is the most well-known example (Devlin et al., 2018). Even though language-specific pre-trained models have been developed for BERT, suggesting poorer language generalizability, the use of models trained on multiple languages is currently being investigated (Pires et al., 2019).

The ultimate aim of the work described in this paper was to find a model that i) can discern between drone and non-drone websites as accurately as possible, ii) has a high recall, and iii) can be applied to webpages written in the languages for the countries studied; e.g. Spain, Ireland, and Italy.

2 Methods

All scripts were developed in Python (v3.7). The countries studied were Spain, Ireland, and Italy. For each country, drone websites were searched for with search words identified in the first stage of the project (GOPA, 2021a); more on that below. Each search engine was searched for with words in both the official language of the country and, for Spain and Italy, also in English. For Spain, for example, a total of 67 combinations of Spanish words and 62 combinations of English words were used, such as the combination of 'drone', 'company' and 'spain'. Appendix A gives an overview of all the word combinations used for Spain. A total of six search engines were used, i.e. Google, Bing, DuckDuckGo, Yahoo, Ask, and AOL. In each case, when available, paid services and the country-specific version of the search engine were used. All links (URLs) returned by each search engine, except those indicated as sponsored, were collected and stored. The total set of URLs obtained – per country – was combined and the domain names were extracted and deduplicated followed by the removal of all domain names with a top-domain country code not indicative of the country studied. After this, a total of 26,067, 14,586, and 53,781 unique domain names were obtained for Spain, Ireland, and Italy, respectively.

Each domain was subsequently scraped. This started with the main page followed by the collection all of other pages – with a maximum of 200 – referred to by those pages within the same domain. The urllib.request (v3.7) function was used to scrape pages. Domains that could not be scraped during the first attempt were visited at least four times – at later points in time – to deal with temporarily unavailable websites. The raw-HTML files obtained were parsed with the Beautiful Soup 4 library (v4.7.1), backed up on the local machine, and processed in several stages. In the first stage, all script and style sections were removed after which the text between the remaining HTML tags was extracted followed by language detection with the langdetect library (v1.07). For Spain and Italy, the official language of the country and English were the only two languages considered; i.e. all non-Spanish or non-Italian web pages were considered written in English. Subsequently, the texts were converted to lower case, and all punctuation marks, numbers, and words below a specific number of characters, either 2 or 3, were removed. Next, depending on the language detected, the words included in the Natural Language Toolkit stop words list (v3.4.1) were removed. No stemming or lemmatization was performed. The texts extracted from all pages collected within the same domain were combined into a single document in which the words were separated by a single space. After processing, a total of 25,829 (99%), 14,476 (99%), and 53,388 (99%) text containing documents remained for Spain, Ireland, and Italy, respectively. From experts, lists of 1,098 Spanish and 686 Italian domain names of drone companies were obtained. These were scraped and processed as described above.

When required, texts were translated by the open source translation software Apertium (2021, v2.0), with the English-Spanish and English-Italian language pairs installed. The creation of a train and test set and model development is discussed in the result section of this paper. The PUlearn (2021, v0.07), scikit-learn (v0.21.2) (Pedregosa et al., 2011), and gensim (v3.4.0) libraries were used for model development. The code and data are available on the GitHub WIH Drones (2022) repository of the project.

3 Results

3.1 Model Development

The initial model was developed on Spanish website texts. The following three classification approaches were compared: 1) Drone word occurrences, 2) Positive and Unlabeled learning, and 3) Supervised Machine Learning.

Table 1: Metrics for the various classification approaches tested.

Type of model	Accuracy	Precision	Recall (TPR)*	TNR	F1
Word-based model	0.76	0.83	0.63	0.88	0.71
PULearn model	0.52	0.16	0.57	0.51	0.25
Logistic Regression, L2 norm	0.85	0.76	0.93	0.79	0.84

*TPR = True Positive Rate, TNR = True Negative Rate

3.1.1 Drone Word-Based

The occurrence of the word drone, ‘dron’ in Spanish, and its most often used abbreviations, i.e. RPAS (Remotely Piloted Aircraft System), UAS (Unmanned Aircraft System), and UAV (Unmanned Aircraft Vehicle), were determined in the text documents obtained. Any document in which at least one of those words occurred, if that word was not a substring of another word, was classified as a drone website. A random sample of 50 drone and 50 non-drone word-containing documents were selected and the files and their corresponding websites were manually inspected by experts. The metrics, i.e. accuracy, precision, recall/True Positive Rate, True Negative Rate, and the F1-score, for the sample are listed in the first row of Table 1.

3.1.2 PUlearning Based

Because a list of 1,098 drone websites was available for Spain and the domain names obtained from the search engines were all unlabeled, a semi-supervised machine learning method known as Positive and Unlabeled Learning (Elkan and Noto, 2008) was applied. This method aims to find the features specific for the set of positive examples and, subsequently, tries to separate the group with those features (the ‘positives’) from the other (the ‘negative’) group in the unlabeled data as good as possible. Apart from the PUlearn algorithm (v0.07), a machine learning classification method able to produce probability estimates is needed. Both Logistic Regression and Support Vector Machine methods can do this and were tested. As only labeled positive examples are known, the challenge when applying this method is to accurately discern the negative and false positive cases. All methods mentioned in Elkan and Noto (2008) to deal with that issue have been implemented in the PUlearn library and were used.

Before training, the documents were converted to a Document Term Matrix (DTM) (Aggarwal, 2016, chap. 13). The DTM was composed of rows, one for each website, and columns that contained the individual words occurring in the entire collection of documents. For each word in the processed texts, the log of the term frequency-inverse document frequency (TF-IDF) + 1 was used as weight (Aggarwal, 2016; Gentzkow et al., 2019). The web page language was added to the DTM as a binary feature (0: Spanish, 1: English), as were the occurrence of particular drone words, i.e. ‘dron’, RPAS, UAS and UAV, in the text and the domain name of the web page visited; eight in total. An 80% random sample was used to train the classifier and the remaining 20% was used as a (holdout) test set.

After comparing all options, a logistic regression model (L2-norm) combined with the PUlearn Elkanto classifier was found to provide the best results with a minimum document frequency of 100 and a minimum character length of 3 for the words included. The ‘accuracy’ of the model on the test set was reported to be 87% according to the library used. However, manual inspection by experts of a randomly drawn sample of 50 drone and 50 non-drone classified cases

Table 2: Metrics on the test set and settings used for various Machine Learning classification algorithms.

Type of algorithm	Accuracy	Precision	Recall (TPR)	TNR	F1	Best (hyper)parameters*
Logistic Regression, L2 norm	0.87	0.88	0.93	0.78	0.90	solver ‘liblinear’, mindf 100
Logistic Regression, L1 norm	0.86	0.86	0.93	0.74	0.90	solver ‘liblinear’, mindf 100
Gradient Boosting classifier	0.86	0.86	0.95	0.70	0.90	estimators 150, mindf 50
Support Vector classifier, rbf kernel	0.85	0.88	0.90	0.76	0.89	gamma 0.3, mindf 200
Support Vector classifier, linear kernel	0.84	0.88	0.89	0.76	0.88	gamma 0.4, mindf 150
Support Vector classifier, poly kernel	0.83	0.82	0.95	0.62	0.88	gamma 0.4, mindf 200
Decision Tree classifier	0.83	0.88	0.88	0.74	0.88	criterion ‘entropy’, mindf 150
Random Forrest classifier	0.83	0.83	0.93	0.66	0.88	estimators 20, mindf 100
Multi-layer Perceptron classifier	0.83	0.87	0.88	0.74	0.88	layers 150, max iter 300, mindf 200
Nearest Neighbors (k = 2)	0.83	0.83	0.93	0.66	0.87	algorithm ‘auto’, p 2, mindf 50
Gaussian Naive Bayes	0.81	0.91	0.79	0.85	0.84	mindf 150
Quadratic Discriminant Analysis	0.80	0.78	0.96	0.51	0.86	mindf 250

*mindf = minimum document frequency, a maximum document frequency of 2000 was used in all cases

gave a completely different view. The actual accuracy was only 52%. The findings are shown in the second row of Table 1.

3.1.3 Supervised Machine Learning based

Next, supervised Machine Learning was applied. It required the additional availability of a set of identified non-drone websites. Since 1,098 positive cases were available, a random selection of about 3,000 domain names from the unlabelled data was manually inspected. After classifying an exact total of 3,059 domains, 2,699 non-drone and 360 drone websites were identified. These were combined with the 1,098 positive cases already available and deduplicated. Next, a DTM was produced exactly as described in the previous section. An 80% random sample was used to train the classifier and the remaining 20% was used as a test set. A whole range of scikitlearn (Pedregosa et al., 2011) classification algorithms was used with (hyper)parameter tuning and 5-fold cross-validation to assure the best possible outcome was obtained for each method (Bergstra et al., 2011). The algorithms, metrics and settings used are shown in Table 2.

It was found that Logistic Regression (L2-norm) provided the best results with an accuracy of 87% on the test set. The settings required to obtain the best results are shown in Table 2 and all used a minimum character length of 3 for the words included. Slightly lower results were obtained for Logistic Regression with an L1-norm (86%), Gradient Boosting (86%), and Support Vector Machine (rbf-kernel; 85%). All other algorithms tried had an accuracy of 84% or lower. Much to our surprise, applying Word Embeddings, either alone (85%) or in combination with the word-frequency based DTM (87%), as described in Daas and van der Doef (2020), followed by Logistic Regression L2 classification, did not result in more accurate results than those based on the DTM alone. This suggested that the simple bag-of-words approach already seemed to extract the most important information from the texts. This and the fact that logistic regression-based findings are very transparent, a requirement essential for application of the findings in official statistics (United Nations, 2014), made us decide to discontinue investigating other Word Embedding or Transformer-based approaches.

From the results of the logistic regression model (L2-norm), a random sample of 50 drone and 50 non-drone classified cases was inspected by experts which resulted in the findings shown in the third row of Table 1. Here, the high recall is important to notice. Comparing the findings in Table 1 reveals that the supervised logistic regression (L2-norm) model developed on the texts extracted from Spanish websites provided the overall best classification results. The model included 2,371 features of which 2,362 were words. The other features were the language and the features indicative of the occurrence of the words 'dron', RPAS, UAV, and UAS in either the document or the domain name. The trained logistic regression model was also able to provide the probability of a text being extracted from a drone website.

3.2 Developing a Generic Model

The training and test set documents were composed of 80% Spanish and 20% English texts. This indicated that the model was trained on a combination of Spanish and English words and suggested a way to improve the classification accuracy of the model. When all Spanish words in the documents were translated into English, with Apertium (2021) software, followed by retraining of the logistic regression L2-model, it was found that the accuracy of the model increased from 87 to 88%. In both cases, these are the averages of 1,000 repeats in which the model was trained on an 80% random sample followed by determining the accuracy on the remaining 20% test set. This finding suggested an interesting way to obtain a model that could – potentially – be applied to the texts extracted from websites in the other countries studied.

However, inspecting the features included in the retrained model revealed that not all words were translated and that some very specific words – related to Spain or to periods in time – were included in the model. The latter could make the model rather specific for Spain and for the period in which the data was collected. Examples of non-translated words are 'españa', 'dron', and 'europa'; apparently, Apertium was unable to translate those words. As a first step, all non-translated words were replaced with their correct English translation in all texts. Examples of words specific to Spain are 'Madrid', 'Barcelona', and other names of cities and regions. Examples of time-period-related words are the months of the year and days of the week. Subsequently, in an iterative approach, all Spanish and time-related features included in the model were removed from the training and test data, followed by retraining. This was repeated until the model did no longer include any Spanish and time-period indicative features. This took a total of six iterations. The model thus obtained still had an accuracy of 88% on the test set (again the average of 1,000 repeats) and contained 1,568 features of which 1,560 were words. The other features were all indicative of the occurrence of the words 'drone', RPAS, UAV, and UAS in

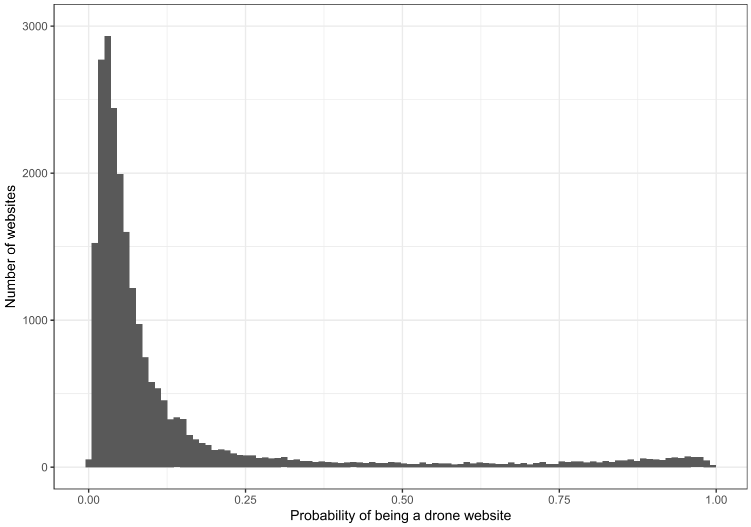

Figure 1: Probability distribution of being a drone website for Spanish websites.

Table 3: Metrics for the Spanish, Irish and Italian samples classified by the same model.

Country	Accuracy	Precision	Recall (TPR)	TNR	F1
Spain	0.86	0.79	0.93	0.80	0.86
Ireland	0.86	0.72	1.00	0.78	0.84
Italy	0.82	0.67	0.97	0.73	0.80

either the document or the domain name; for the latter case, the language-specific word 'dron' was used. The language feature was no longer included as all texts were written in the same language.

Applying the model to all text extracted from Spanish websites, after translating all words to English (including those identified above), produced 25,829 results with the probability distribution shown in Figure 1. Here, a clear distinction between a very large group of obviously non-drone websites and a small group of potential drone websites can be seen. When a cut-off value of 0.5 is used to identify drone websites, a total of 2,139 websites were found (8.3%). To check these findings, random samples of 50 were drawn from 10 probability ranges, each 0.1 wide, and the corresponding websites were manually checked by experts. For a cut-off value of 0.5, this resulted in the metrics shown in the first row of Table 3. It revealed an accuracy of 86%, which is somewhat lower than the value reported on the test set but higher than the value in row three of Table 1. Also a high recall of 93% was found.

From the distribution in Figure 1 it is clear that very large numbers of obvious non-drone websites occur in the data set. This indicates that drone websites are the minority class. A fairly broad, slowly increasing number of websites can be observed at probabilities values of 0.5 and higher. This may indicate that different types of drone websites exist and could also suggest that 0.5 may not be the best cut-off value for drone website detection. Results for different cut-off values were compared and indicated a major downside when applying these. For instance, for a cut-off value of 0.6, the accuracy, precision, and recall were 85%, 83%, and 81%, respectively. Comparing those findings with the ones shown in the first row of Table 3, show an increase in

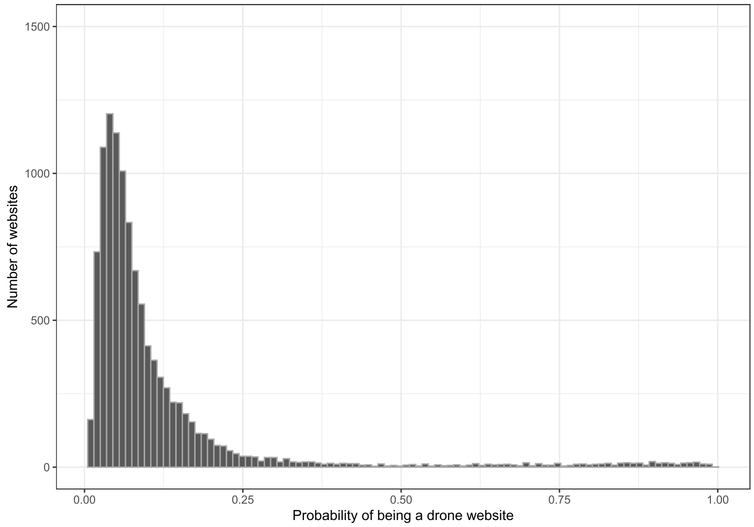

Figure 2: Probability distribution of being a drone website for Irish websites.

the precision but a rapid decrease in recall. After these positive classification results, the model was applied to the texts extracted from Irish websites.

3.2.1 Application to Irish Websites

First, the language of the texts extracted from the Irish websites was checked in more detail as the texts could potentially be written in English and Irish (Gaeilge). Here, the langid (v1.16) library was used as this is the python library with the highest recall on identifying Irish texts; see the recall table on Fasttext (2022). These results confirmed that all texts were indeed written in English; none were identified as Irish. The texts were subsequently converted to a DTM exactly as described above to which features were added indicative of the occurrence of the words ‘drone’, RPAS, UAV, and UAS in the document and the domain name, respectively. Next, the records in the DTM were classified with the model.

Classification produced a total of 14,476 probability estimates. Assuming a cut-off value of 0.5, a total of 785 drone websites (5.4%) were found. The distribution of the probabilities is shown in Figure 2. Here, a very similar distribution to the one shown in Figure 1 is observed; although the right part seems to be somewhat flatter. To validate the results obtained, random samples of 50 were drawn from each of the 10 bins discerned and the corresponding websites were manually checked by experts. The metrics of these findings are shown in the second row of Table 3. They are very similar to those found for the Spanish websites, except for a somewhat lower precision. The perfect recall is worth mentioning. Results for different cut-off values were compared and revealed a considerable reduction in recall for higher values. For instance, a cut-off value of 0.6 resulted in an accuracy, precision, and recall of 84%, 75%, and 83%, respectively. When all results are taken into consideration, it can be concluded that the Spanish website based model is indeed able to identify drone websites in another country. This led the way to the development of an approach that could be used to identify Italian drone websites.

3.2.2 Application to Italian Websites

To enable the classification of Italian websites, with the model developed, three steps needed to be performed. The first was creating a list of Italian words, analog to the original translated Spanish words included (as features) in the model; i.e. a total of 1,560 words. We will refer to this list from here on as the 'translation list'. The list is created by translating the English words included in the model into Italian but (and this is important) is used – in practice – to translate Italian words into English. The second step is translating the occurrences of those Italian words – and nothing else – in the texts extracted from all Italian websites. This and the addition of the features indicative of the words 'drone', RPAS, UAV, and UAS in the document and the domain name, will enable the classification of Italian websites by the model. The third and last step is to validate the classification findings obtained. In this process, the list of 686 Italian drone websites plays an important role.

First, the 1,560 English words included as features in the model were translated into Italian with Apertium and the Italian-English module. Any non-translated words were, subsequently, translated with Google Translate (English to Italian). All translations were manually checked by the authors and a native speaker of Italian. Here, it became clear that in some cases the translation was incomplete, not needed, or incorrect. Incomplete translation occurred for words with a feminine and masculine variant in Italian. For example, 'one' is the English translation of 'una' (female) and 'uno' (male); so both cases need to be included in the translation list. A total of sixteen words required this. In four cases, it was found that translation was not required as the English words are commonly observed on Italian (and Spanish) web pages. These words are 'web', 'cookie', 'cookies', and 'log'. Hence, these words were removed from the translation list. Two words were found that were incorrectly translated; namely, 'fly' and 'unmanned'. Translating those words in Italian, in the context of the drone industry, should result in 'volare' and 'senza pilota', respectively. The automatic translation procedure, however, suggested 'mosca' and 'senza equipaggio'. These incorrect translations were replaced with their correct ones. In the end, a list composed of 1,572 Italian to English word translations was created.

To verify if the combination of the translation list and trained model produced valid results, it was first tested on the documents obtained from the list of 686 Italian drone companies. After translating the Italian words included in the list to English for all texts and converting them to a DTM, as described for the Irish and Spanish documents, the model identified 584 documents as a drone company; i.e. it had an accuracy of 85%. This is very similar to those shown in Table 3 for the Spanish and Irish documents. Any additional changes made to the translation file, with the intention to improve it, never increased the accuracy of classification; on the contrary, it usually decreased to values of 75% or lower. This indicated that the creation of a translation list is a very delicate task. In our experience, it should be kept as simple as possible. Subsequently, the model was applied to all the 53,388 Italian text documents after translating and processing as described above. This resulted in 53,388 probability values for which, for a cut-off value of 0.5, a total of 2,052 (3.8%) were identified as drone websites.

The distribution of the probability values for all classified cases is shown in Figure 3. This distribution differs from those shown in Figures 1 and 2. Especially in the low probability range, below values of 0.5, the Italian results display a much broader distribution. For each of the 10 bins discerned random samples of 50 were drawn and the corresponding websites were manually checked by experts. The metrics of these findings are shown in the third row of Table 3. For most of them, the values are somewhat lower than those observed for the Spanish and Irish websites, indicating a small, but noticeable, lower performance. The recall, however, was again very high

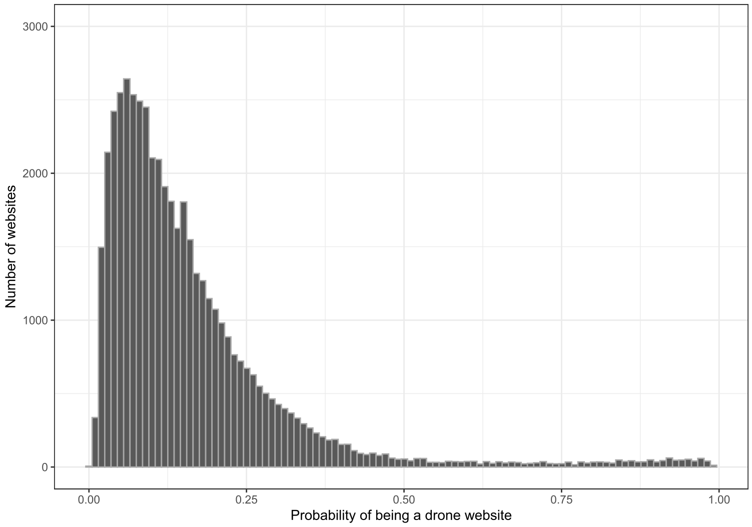

Figure 3: Probability distribution of being a drone website for Italian websites.

in this case. Comparing the results for different cut-off values again showed a negative effect on the recall. For instance, a cut-off value of 0.6 resulted in an accuracy, precision, and recall of 84%, 74%, and 85%, respectively. All findings indicate the Spanish website-based model is able to identify Italian drone websites, but with a slightly lower accuracy. Based on the recall, however, the model is very well suited to select potential drone company websites.

4 Discussion

The findings described in this paper reveal that its possible to discern drone from non-drone websites based on their texts. A supervised logistic regression model with an L2-norm performed best when trained on Spanish websites when all Spanish words were translated into English. Manual checking of a sample revealed an accuracy of 86% and a recall of 93% (Table 3). Hyperparameter optimization and/or including word embedding-based features did not produce better performing models suggesting that the model already performed close to what maximally could be obtained.

Inspecting the features included in the model revealed that a considerable number of those features were very specific for Spain or indicative of certain moments in time. When these features were removed from the texts – followed by retraining the model – a completely English word-containing model was obtained with identical metrics. That model could be successfully applied to identify websites of drone companies in Ireland and Italy. This suggested that the model developed had successfully captured the concepts associated with the words used in each of the countries studied.

Classification of Italian websites required a translation list of Italian to English words. Although the performance of the model on translated Italian websites was slightly lower than those found for Irish and Spanish websites, all results had a very high recall in common; between 93 and 100% (see Table 3). This shows that the model is particularly well suited for the identification of (potential) drone websites in large collections of websites (for the three countries studied). The low precision of the model, however, indicates that these selected, positively classified, websites will certainly contain considerable amounts of false positives; i.e. non-drone websites classified as drone websites. This does not have to be a downside as this is exactly the task for which the model was used in our web-oriented drone study (Daas et al., 2022). With the model, large amounts of potentially interesting websites for each of the countries studied were processed after which a total of 2,139 (8.3%), 785 (5.4%), and 2,052 (3,8%) remained for Spain, Ireland, and Italy, respectively. The percentages between brackets indicate the effectiveness of this approach. Subsequent detailed analysis of the small number of websites remaining needed to be performed to remove the false positive from the actual drone websites. This included manual checking (Daas et al., 2022).

In future studies, it would be interesting to verify if the creation of translation lists for other languages, such as Dutch, German or French, could – in combination with the model – produce results as reliable as those obtained for the identification of Italian drone websites; especially regarding its high recall. Such a study would greatly benefit from the availability of lists of drone company websites for each of those countries. We found that such a list greatly assisted the validation of the accuracy of the translation list produced. However, this is not needed for websites written in English. Preliminary results reveal that the model is indeed also able to correctly identify drone websites written in English for countries such as the Netherlands.

A Appendix: The Search Word Combinations Used for Spain

Spanish Word Combinations: dron empresa espana, rpas empresa espana, uav empresa espana, uas empresa espana, dron negocio espana, rpas negocio espana, uav negocio espana, uas negocio espana, dron tienda espana, rpas tienda espana, uav tienda espana, uas tienda espana, dron curso formacion espana, rpas curso formacion espana, uav curso formacion espana, uas curso formacion espana, dron piloto espana, rpas piloto espana, uav piloto espana, uas piloto espana, dron operador espana, rpas operador espana, uav operador espana, uas operador espana, dron proveedor servicios espana, rpas proveedor servicios espana, uav proveedor servicios espana, uas proveedor servicios espana, dron fabricante espana, rpas fabricante espana, uav fabricante espana, uas fabricante espana, dron u-space espana, rpas u-space espana, uav u-space espana, uas u-space espana, dron componentes espana, rpas componentes espana, uav componentes espana, uas componentes espana, dron accesorios espana, rpas accesorios espana, uav accesorios espana, uas accesorios espana, dron software espana, rpas software espana, uav software espana, uas software espana, (dron OR rpas OR uav OR uas) socio espana, (dron OR rpas OR uav OR uas) miembro espana, (dron OR rpas OR uav OR uas) registro espana, (dron OR rpas OR uav OR uas) inscripcion espana, (dron OR rpas OR uav OR uas) asociacion espana, (dron OR rpas OR uav OR uas) comunidad espana, (dron OR rpas OR uav OR uas) red espana, (dron OR rpas OR uav OR uas) descripcion espana, (dron OR rpas OR uav OR uas) lista espana, (dron OR rpas OR uav OR uas) socio espana pdf, (dron OR rpas OR uav OR uas) miembro espana pdf, (dron OR rpas OR uav OR uas) registro espana pdf, (dron OR rpas OR uav OR uas) inscripcion espana pdf, (dron OR rpas OR uav OR uas) asociacion espana pdf, (dron OR rpas OR uav OR uas) comunidad espana pdf, (dron OR rpas OR uav OR uas) red espana pdf, (dron OR rpas OR uav OR uas) descripcion espana pdf, (dron OR rpas OR uav OR uas) lista espana pdf

English Word Combinations: drone company spain, rpas company spain, uav company spain, uas company spain, drone business spain, rpas business spain, uav business spain, uas business spain, drone shop spain, rpas shop spain, uav shop spain, uas shop spain, drone training course spain, rpas training course spain, uav training course spain, uas training course spain, drone pilot spain, rpas pilot spain, uav pilot spain, uas pilot spain, drone operator spain, rpas operator spain, uav operator spain, uas operator spain, drone service provider spain, rpas service provider spain, uav service provider spain, uas service provider spain, drone manufacturer spain, rpas manufacturer spain, uav manufacturer spain, uas manufacturer spain, drone u-space spain, rpas u-space spain, uav u-space spain, uas u-space spain, drone components spain, rpas components spain, uav components spain, uas components spain, drone accessories spain, rpas accessories spain, uav accessories spain, uas accessories spain, drone software spain, rpas software spain, uav software spain, uas software spain, (drone OR rpas OR uav OR uas) members spain, (drone OR rpas OR uav OR uas) register spain, (drone OR rpas OR uav OR uas) registration spain, (drone OR rpas OR uav OR uas) association spain, (drone OR rpas OR uav OR uas) community spain, (drone OR rpas OR uav OR uas) overview spain, (drone OR rpas OR uav OR uas) list spain, (drone OR rpas OR uav OR uas) members spain pdf, (drone OR rpas OR uav OR uas) register spain pdf, (drone OR rpas OR uav OR uas) registration spain pdf, (drone OR rpas OR uav OR uas) association spain pdf, (drone OR rpas OR uav OR uas) community spain pdf, (drone OR rpas OR uav OR uas) overview spain pdf, (drone OR rpas OR uav OR uas) list spain pdf

Acknowledgement

We thank the employees of Eurostat involved in the project for their helpful comments on earlier drafts of the reports on which part of the work described in this paper is based. We especially thank Florabella Carausu (GOPA) and Virginia Santamarina (Univ. Politècnica de València) for their guidance, constructive comments, and assistance in the manually checking of Spanish and Italian drone websites. Andrea Ascheri (Eurostat) is gratefully acknowledged for carefully verifying the Italian-English translation list. We also extend our acknowledgments to Paola Olivares (Politecnico di Milano) for sharing the list of Italian drone company websites.

Funding

This research was performed as part of the study "Web intelligence for measuring emerging economic trends: the drone industry" led by GOPA under the framework contract on Methodological Support (Ref. 2018.0086) for Eurostat.

References

Aggarwal C (2016). *Data Mining: The Textbook*. Springer, New York.

Almeida F, Xexéo G (2019). Word embeddings: A survey. *CoRR*, arXiv preprint: https://arxiv.org/abs/1901.09069

Antonacopoulos A, Hu J (2003). *Web document analysis: Challenges and opportunities*. World Scientific Publishing Co. Pte. Ltd., Singapore.

Apertium (2021). Website of apertium, a free/open-source machine translation platform. http://www.apertium.org.

Aweisi A, Arora D, Emby R, Rehman M, Tanev G, Tanev S (2021). Using web text analytics to categorize the business focus of innovative digital health companies. *Technology Innovation Management Review*, 11(7/8): 65–78.

Bergstra J, Bardenet R, Bengio Y, Kégl B (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., New York.

Beręsewicz M, Pater R (2021). *Inferring job vacancies from online job advertisements. Statistical Working papers*. Eurostat, Luxembourg.

Daas P, de Wolf N (2021). Identifying different types of companies via their website text. In: *Symposium on Data Science and Statistics (SDSS)*. Virtual, June 2-4, 2021.

Daas P, Tennekes M, De Miguel B, De Miguel M, Santamarina V, Carausu F (2022). *Web intelligence for measuring emerging economic trends: The drone industry Statistical Working papers*. Eurostat, Luxembourg.

Daas P, van der Doef S (2020). Detecting innovative companies via their website. *Statistical Journal of IAOS*, 36(4): 1239–1251.

De Kunder M (2022). The size of the world wide web (the internet). https://www.worldwidewebsize.com/.

Devlin J, Chang MW, Lee K, Toutanova K (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint: https://arxiv.org/abs/1810.04805, 13 pages.

Elkan C, Noto K (2008). Learning classifiers from only positive and unlabeled data. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Y Li, B Liu, S Sarawagi, eds.). Las Vegas, Nevada, USA. August 24–27, 2008, 213–220. ACM.

ESSnet (2020). Web page provinding an overview of the experimental statistics developed in the context of essnet big data workpackage C on enterprise characteristics. https://ec.europa.eu/eurostat/cros/content/wpc-experimental-statistics_en.

Fasttext (2022). Webpage of fasttext language detect v1.0.3. https://pypi.org/project/fasttext-langdetect/.

Florescu D, Karlberg M, Reis F, Rey Del Castillo P, Skaliotis M, Wirthmann A (2014). Will 'big data' transform official statistics? Quality in Official Statistics Conference. Vienna, Austria. June 2-5, 2014.

Gentzkow M, Kelly B, Taddy M (2019). Text as data. *Journal of Economic Literature*, 57(3): 535–574.

GitHub WIH Drones (2022). Web intelligence hub drone companies. https://github.com/eurostat/wih_drones_companies.

Gökk A, Waterworth A, Shapira P (2015). Use of web mining in studying innovation. *Scientometrics*, 102(1): 653–671.

GOPA (2021a). Data Retrieval, Deliverable 2. Report 2 of the project Web Intelligence for Measuring Emerging Economic Trends: The Drone Industry. Eurostat, Luxembourg.

GOPA (2021b). Deliverable 1. Report 1 of the project Web Intelligence for Measuring Emerging Economic Trends: The Drone Industry. Eurostat, Luxembourg.

Höchtl J, Parycek P, Schöllhammer R (2015). Big data in the policy cycle: Policy decision making in the digital era. *J. Org. Comp. Elec. Com.*, 26(1–2): 147–169.

Kitchin R (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS*, 31(3): 471–481.

Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D (2019). Text classification algorithms: A survey. *Information*, 10(4).

Kühnemann H, van Delden A, Windmeijer D (2020). Exploring a knowledge-based approach to predicting nace codes of enterprises based on web page texts. *Statistical Journal of the IAOS*, 36(3): 807–821.

Larose D, Markov Z (2007). *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*. Wiley-Interscience, Hoboken, NJ.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pires T, Schlinger E, Garrette D (2019). How multilingual is multilingual BERT? In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4996–5001. Association for Computational Linguistics, Florence, Italy.

Powell B, Nason G, Elliott D, Mayhew M, Davies J, Winton J (2018). Tracking and modelling prices using web-scraped price microdata: Towards automated daily consumer price index forecasting. *Journal of the Royal Statistical Society: Series A*, 181(3): 737–756.

PUlearn (2021). Website of the pulearn python library v0.07. https://pypi.org/project/pulearn.

Rothaermel F (2019). *Strategic Management*. McGraw-Hill Education, New York.

Song M, Wu YF (2008). *Handbook of Research on Text and Web Mining Technologies*. Information Science Reference, Hershey, NY.

United Nations (2014). Fundamental Principles of Official Statistics. United Nations Statistic Division, New York.