applied sciences

MDPI

Article

Identifying Polarity in Tweets from an Imbalanced Dataset about Diseases and Vaccines Using a Meta-Model Based on Machine Learning Techniques

Alejandro Rodríguez-González [1,2,*], Juan Manuel Tuñas [1], Lucia Prieto Santamaría [1], Diego Fernández Peces-Barba [1], Ernestina Menasalvas Ruiz [1,2], Almudena Jaramillo [3], Manuel Cotarelo [3], Antonio J. Conejo Fernández [4], Amalia Arce [5] and Angel Gil [6]

1 Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarcón, Madrid, Spain; juan.tunas@ctb.upm.es (J.M.T.); lucia.prieto@ctb.upm.es (L.P.S.); diego.fernandez@ctb.upm.es (D.F.P.-B.); ernestina.menasalvas@upm.es (E.M.R.)
2 Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain
3 Global Medical and Scientific Affairs, MSD España, 28027 Madrid, Spain; almudena_jaramillo@merck.com (A.J.); manuel.cotarelo@merck.com (M.C.)
4 Hospital Vithas Xanit Internacional, Benalmádena, 29630 Málaga, Spain; ajconejof@gmail.com
5 HM Nens, Barcelona, 08009 Cataluña, Spain; aarce@hmhospitales.com
6 Departamento de Especialidades Médicas y Salud Pública, Facultad de Ciencias de la Salud, Campus de Alcorcón, Universidad Rey Juan Carlos, 28933 Madrid, Spain; angel.gil@urjc.es
* Correspondence: alejandro.rg@upm.es

Received: 16 November 2020; Accepted: 14 December 2020; Published: 17 December 2020

Abstract: Sentiment analysis is one of the hottest topics in the area of natural language. It has attracted a huge interest from both the scientific and industrial perspective. Identifying the sentiment expressed in a piece of textual information is a challenging task that several commercial tools have tried to address. In our aim of capturing the sentiment expressed in a set of tweets retrieved for a study about vaccines and diseases during the period 2015–2018, we found that some of the main commercial tools did not allow an accurate identification of the sentiment expressed in a tweet. For this reason, we aimed to create a meta-model which used the results of the commercial tools to improve the results of the tools individually. As part of this research, we had to deal with the problem of unbalanced data. This paper presents the main results in creating a metal-model from three commercial tools to the correct identification of sentiment in tweets by using different machine-learning techniques and methods and dealing with the unbalanced data problem.

Keywords: machine learning; unbalanced data; metamodel; sentiment analysis; commercial tools; twitter data

1. Introduction

Nowadays, information can be obtained from a vast number of sources, most of which are available on the Internet. It is well known that the Internet has become our main knowledge engine, redefining the way we communicate and gain understanding about the world around us. That implies that information of every kind and from every field can be publicly accessed just by typing a few words into our navigator.

Among the wide range of knowledge areas, searches for information on the Internet related to healthcare are common. People can solve health doubts and find data in a quick and easy manner. If a certain health issue concerns the population, such searches tend to increase. In the past, some studies

about these web queries have corroborated that fact [1–3] and have been used in the pursuit of improving public healthcare strategies and social wellbeing. Some of such health issues can be related to the early detection of disease outbreaks [4–7], disease surveillance [8–11] or epidemic intelligence [5,12] and even in the current COVID-19 outbreak, infoveillance studies can aid in tackling the pandemic situation [13–15]. However, the diseases that have been screened under this type of analyses are not limited to COVID-19 as it will be discussed later.

The Internet has allowed the emergence of new services and applications. Some of the most popular and with an increasing use, can be grouped under the term "social media". Social media is a subject of study for a broad variety of domains related to computer science, since it can be seen as an important means of data to analyze users' feelings and opinions. Those approaches that have taken advantage of the information present in social media have traditionally included political [16] and marketing campaigns [17–20] or financial predictions [21–23]. Lately, it has been proven that in the health scope, social media information can be relevant too by assessing epidemiological patterns [24], predicting epidemic outbreaks [25,26] or detecting drug side effects [27] and medication safety [28]. Moreover, concerns about vaccines and vaccination are widely expressed in social media [29,30]. With the expansion of anti-vaccines movements in the recent years, the debate in social media has only grown and several conversations about vaccines have been monitored [31–34]. Special controversy has surrounded the human papillomavirus (HPV) vaccines [29,35–43]. Besides, some of the diseases that have been studied under these types of approaches of Internet and social media mining comprise largely Influenza [4,5,9,11,25,44–46] but also others such as Zika virus disease [47,48], cholera [24], obesity [49] or diabetes [50].

Twitter, amongst all the possibilities in social media, has been explored in numerous works, as it is the perfect platform to share opinions that can be mined [17,19,22,23]. Sentiment analysis or opinion mining stands for natural language processing (NLP) methods that aim to computationally extract, interpret and classify emotions and subjective information from unstructured resources. The attitude of authors towards a topic in a text can be categorized as being positive, negative or neutral. In other words, we can assign a polarity to a text. The applications of these methods involve multiple domains (political science, social sciences, market research, etc.) and have evolved over time [51]. Nowadays, the most important sources to mine in this context come from the Internet and, as it has already been mentioned, Twitter can be seen as one of the most popular. The research on sentiment analysis in Twitter is called Tweet Sentiment Classification (TSC) [52] and multiple works have been developed under this approach [19,20,22,23,53]. One important topic that has attracted attention on Twitter has been vaccination. Analysis of discussions, opinions and feelings about certain vaccines in tweets has been performed to detect the feelings related to vaccine promotion [36,39–41].

In the current paper, we analyze the possibilities of exploring social media information, and in particular Twitter, in order to extract feelings related to different vaccines messages. Thus, the expansion of negative opinions related to a set of vaccines and their related diseases could be monitored. The analysis has been performed by mining tweets in Spanish language published during the period 2015–2018. We have created several classification meta-models according to different machine learning techniques and different datasets. As collected data were imbalanced, sampling methods were used to address the situation. The work is included in the MAVIS study and it is an extension of a previous work [54]. The structure of the present paper consists of the following sections: Section 2 includes the methodology performed, Section 3 details the obtained results, while Section 4 discusses them. Finally, Section 5 summarizes the achieved conclusions and the future work to be carried out.

2. Materials and Methods

In this section, the pipeline to gather the data from Twitter, perform the sentiment analysis and create classification models is detailed. Some of the processes here explained were previously presented in [54]. The first objective was, using a set of vaccines and their related diseases, to discover whether a negative opinion about them was spreading in Twitter or not. To this aim, first Twitter messages

associated to those concepts were extracted. Afterwards, the tweets' polarity was classified both by three commercial tools and by five evaluators, who annotated the Tweets manually. As the class assigned to tweets was distributed in an imbalanced manner, sampling methods were performed to obtain different datasets. To end with, several machine learning techniques were applied to data to generate different classification models.

2.1. Twitter Data Extraction and Sentiment Analysis

The keywords on which this study has focused its interest are related to the following vaccines and diseases:

- Invasive meningococcal disease ("EMI" in Spanish): Bexsero, Trumenba, Nimenrix.
- Invasive pneumococcal disease ("ENI" in Spanish).
- Influenza.
- Hepatitis.
- Rotavirus: Rotarix, Rotateq.
- Measles ("Sarampión" in Spanish) and MMR ("Triple vírica" in Spanish).
- Sepsis.
- Whooping cough ("Tosferina" in Spanish).
- Chickenpox ("Varicela" in Spanish): Varivax, Varilrix; and Shingles ("Zoster" in Spanish).
- Human papillomavirus infection ("VPH" in Spanish): Cervarix, Gardasil.

Although Instagram data were also considered for the current methodology, the amount of data compared to Twitter was very low so for the sake of quality models generation, Instagram data were discarded.

Twitter data were obtained by using the official API (Application Programming Interface), from which all the required information for the current study could be extracted. The execution of the extraction process obtained a total of 1,028,742 tweets, from which 318,302 were different/original tweets. The number of retweets was 10,440 and the number of quotes of the original tweets was 65,806. The keywords used in the search were mentioned 1,187,046 times. After extracting all the tweets, they were all submitted to a cleaning process in order to get a consistent and understandable version of the texts. Hashtags (#), user mentions (@), URLs, email addresses, retweet markers (RT:) and emojis and other non-representable characters were removed.

Sentiment analysis examines the content of free-text natural language to identify opinions and emotions. One of the principals and most important sources of text comes from the Internet and it is social media. Sentiment analysis from social media has been a common topic on research and diverse software tools have been developed to automatize its processes, enabling the classification of large numbers of texts [51]. Methods may focus on the polarity of texts ("positive", "negative", "neutral") but also can be centered on feelings and emotions ("angry", "happy", "sad") or intentions ("interested", "not interested"). Sentiment analysis approaches have been categorized in three main groups: knowledge-based, statistical and hybrid methods [55]. Efforts have been made to extract sentiments associated with polarities of positive or negative for specific subject of a text, instead of classifying the whole text as positive or negative [56].

2.2. Different Datasets Creation

To get the different datasets that would be input to the machine learning methods to generate the classification models, a process to annotate the tweets was implemented. Such annotations were performed in two ways: (1) using different commercial tools and (2) being manually revised by expert evaluators.

- Annotation with commercial tools

Three tools were chosen to automatically annotate the whole set of tweets and quotations: IBM Watson (https://www.ibm.com/watson/services/tone-analyzer/) (now called Watson Tone Analyzer), Google Cloud Natural Language (https://cloud.google.com/natural-language) and Meaning Cloud (https://www.meaningcloud.com/es). The three of them returned different formats of the polarity of tweets. IBM and Google returned numerical values (a score between −1 and 1) while Meaning Cloud returned a class (a class value between 6 classes: P+, P, NEU, N, N+ and NONE). The analysis was simplified so that tweets were discretized to either "negative" or "non-negative". This way, "non-negative" would include other classes such as "neutral". Models were generated considering (i) the original values without discretizing ("original"), (ii) the adapted discretized values ("adapted") and (iii) both of them ("both").

- Manual annotation by experts

Five experts in the field classified the tweets' polarity manually, by determining if each tweet contained a "negative" or a "non-negative opinion". An iterative process of annotation was cyclically performed three times: in each iteration a set of 100 tweets were annotated classifying the sentiment expressed in tweets. A total of 300 tweets were annotated. From those, a very low number of tweets were identified as negative, leading to a class imbalance challenge that could impact on the quality of the models that would be later generated.

To increase the number of negative tweets, a sample that contained words with negative sentiments was extracted from the original dataset. Such words were selected from Meaning Cloud platform, since it allows retrieving the polarity of the words extracted in the sentiment analysis process. A subset of 459 tweets containing words from this list was obtained, ensuring they had not been previously selected for the first three iterations. Therefore, the total number of tweets that were classified by the five evaluators amounted to 759.

Analyzing the resulting manually annotated dataset, it was found that two of the experts had a high grade of disagreement with the other three evaluators. Annotations of those three and of the total five evaluators were compared to the annotations performed by the commercial tools. The level of agreement was higher between the three evaluators and the commercial tools than between the five evaluators and the commercial tools. For the three evaluators, there were 142 tweets classified as negative and 617 as non-negative; and, for the three evaluators, there were 128 tweets annotated as negative and 631 as non-negative. Nevertheless, as will be stated in the next subsection, models were generated with both the three and the five expert's annotations, as the difference between them was sparse and the five evaluators scenario increased the number of negative tweets.

2.3. Sampling and Models Generation

Learning from imbalance data is still a challenge for machine learning methods nowadays [57]. A classifier, when trained with an imbalance distribution dataset, tends to be biased towards the more frequent class. Therefore, efforts to avoid such types of skewed distributions are of paramount importance to ensure the aforementioned influence is not learned by the model. Pre-processing or training methods should then focus on alleviating this disadvantage. The aim is to create a learning system that is able to predict over the minority class but without sacrificing the performance on the majority one [58].

As this is not a trivial challenge and has a major relevance in the generated model accuracy, multiple approaches to solve it have been proposed in the literature [59,60]. These types of approaches are called sampling methods. There are two main ways of balancing an imbalanced class set: downsizing the large class (also known as under-sampling) or upsizing the small class (also known as over-sampling). Generally, over-sampling is preferred [61]. Both can be performed on a random basis by randomly adding or dropping instances of the minority or majority class, respectively. However, this approach may cause overfitting or loss of information.

Thus, other techniques aim to overcome such limitations. One of them [62] introduced a cluster-based under-sampling approach, where it was proposed that clusters in the dataset that have more majority class samples and less minority class samples will behave like the majority class samples, and vice versa. Therefore, it would be reasonable to select a suitable number of majority class samples from each cluster by considering the ratio of the number of majority class samples to the number of minority class samples in the cluster. Regarding over-sampling, SMOTE [63] (synthetic minority over-sampling technique) was developed to generate synthetic minority class examples by selecting new samples close to the existing ones in the feature space. On the other hand, ADASYN [64] (adaptative synthetic sampling) also generates synthetic sample points for the minority class but considers a density distribution to decide the number of synthetic samples to be generated for a particular point.

The main objective of the current research work was to generate different classification models by means supervised machine learning techniques. Such models were obtained starting from the different datasets mentioned in the previous subsection: considering the different numbers of evaluators (3 or 5) and discretizing or not the inputs (original, adapted or both). The variables in such datasets were the following: the output values from (i) IBM (originally a numerical score), (ii) Google Cloud (originally a numerical score) and (iii) Meaning Cloud (originally a discrete class) tools, and the (iv) manually annotated class.

On the other hand, different sampling techniques were performed in order to balance the number of negative and non-negative classes associated to tweets. As discussed in Section 2, imbalanced datasets lead to poor quality classification models. There are two main ways of addressing this challenge: under- or down-sampling (i.e., reducing the number of samples of the majority class) and over- or up-sampling (i.e., increasing the number of samples of the majority class). For down-sampling, two methods were performed: random sampling and clustering; and for up-sampling, three methods were implemented: random sampling, SMOTE and ADASYN (see Section 2 for detailed references). In both cases of random sampling, the method was iterated over 10 times to ensure the randomness. For clustering, k-means was implemented. Both clustering and ADASYN were only applied to the original data not discretized, as they cannot handle categorized input variables.

There is a wide variety of machine learning methods that are used to generate classification models. This kind of machine learning is also known as supervised machine learning. A classifier is a system that can predict, based on the previous learning, the class of a new input instance. Some of the methods that have been discussed in the literature and that will be used for our work are the following: C5.0, Logit Boost, Bayesian Generalized Linear Models (BayesGLM), Multilayer Perceptron, Random Forests (RF) and Support Vector Machine models (SVM).

C5.0 [65] (p. 5) is an improvement of the classic C4.5 algorithm [66], which generates decision trees based on the concept of information entropy. LogitBoost [67], also known as additive logistic regression, applies a boosting approach to build a logit model using decision trees as weak learners [68]. BayesGLM [69] uses an approximate Expectation-Maximization (EM) to fit GLM with the Student-t prior to distribution. The Multilayer Perceptron [70,71] is a class of artificial neural network that uses hidden layers and the back-propagation error algorithm. Random Forests [72] is a metaclassifier that builds multiple decision trees and combines their outputs by a voting process. Finally, SVM [73] build models as function estimation and optimization problems, in a linear or non-linear way, separating classes by hyperplanes.

Some of the literature works related to health sentiment analysis have used machine learning methods such as SVM [29,39], Naïve-Bayes [74], Random Forests and Random Decision Trees [48].

Six supervised learning algorithms were implemented to obtain the models: C5.0, Logit Boost, Bayes GLM, Multilayer Perceptron, Random Forest and SVM. The hyperparameters of each algorithm were either set in their default values or optimized between 4 and 12 parametrizations, for further indications code and packages are provided as it is indicated below. The Multilayer Perceptron used a Weighted Decay in the los function and SVM was implemented with a linear kernel. Each of them

was evaluated using 10-fold cross-validation and their accuracies were computed by the mean and standard deviation of ROC values.

All the sampling and classification methods were implemented in R, mainly by UBL (https://www.rdocumentation.org/packages/UBL) and caret (https://www.rdocumentation.org/packages/caret) packages, respectively. All the code is included in the Supplementary Materials.

3. Results

The principal objective of the present work is the generation of sentiment classification models learned from annotated vaccine-related tweets. Six different supervised learning methods were used for this task: C5.0, Logit Boost, Bayes GLM, Multilayer Perceptron, Random Forests and SVM. Each algorithm was run in each of the sampling subsets (non-sampling, down-sampling and up-sampling) and considering (i) the different numbers of evaluators (3 or 5) and (ii) the distinct input predictors (original, adapted or both). In the random modalities of the up-sampling and down-sampling, the techniques were executed a total of 10 times to analyze the independent results to guarantee their significance.

The full tables with all the mean values and standard deviations of the ROC curves derived from 10-fold cross validation of the previous analysis are included in Supplementary Materials (https://medal.ctb.upm.es/internal/gitlab/mavis/mavis/blob/master/SA_ASC/MAVIS_tables_by_sampling.xlsx). In the provided .xlsx file, each sheet corresponds to each sampling method ("NO sampling", "DOWN–random", "DOWN–clustering", "UP–random", "UP–smote" and "UP–ADASYN"). All the results are shown for every Machine Learning (ML) method and every combination of the number of evaluators and predictors.

For the highest values of ROC curves' mean in each sampling subset and obtained by each ML method, results have been visualized (https://medal.ctb.upm.es/internal/gitlab/mavis/mavis/blob/master/SA_ASC/MAVIS_SA_best_results.xlsx). We have represented three figures: Figure 1 represents the highest accuracy of the models in the initial subset without performing sampling, Figure 2 in the two under-sampling subsets and Figure 3 in the three over-sampling subsets. The colors of the bars stand for the different classification methods and the textures represent the input predictor from which the model has been generated. All the best generated models have been obtained from either the original predictor ("ORIG") or the combination of the original and the adapted ("BOTH"), but none of the best models in any subset have been generated from the adapted predictor ("ADAP"). There were two cases (1 for non-sampling subset and 1 for up-sampling) in which the highest mean of ROC values was the same for both predictors. In those cases, for simplicity, the predictor has been represented as the original one.

The width of the bars in Figure 1 represents the number of evaluators, being 3 in the thinner bars and 5 in the larger one. For the other two figures, all the bars are representing models coming from three evaluators annotations. There were five cases (1 for non-sampling subset, 2 for down-clustering and 2 for up-sampling) in which the highest mean of ROC values was the same for both numbers of evaluators. For the sake of clarity in those cases, the number of evaluators has been set to three.

Overall, the result of averaging the accuracy mean of the generated models with the ML methods in the different sampling subsets, led to the results presented in Table 1. In this table, the mean values of the different ML methods' accuracies from each of the sets have been represented. Results are displayed ranking the applied ML tools from left to right so that the higher global accuracy averages are shown in the left while the lower ones appear to the right.

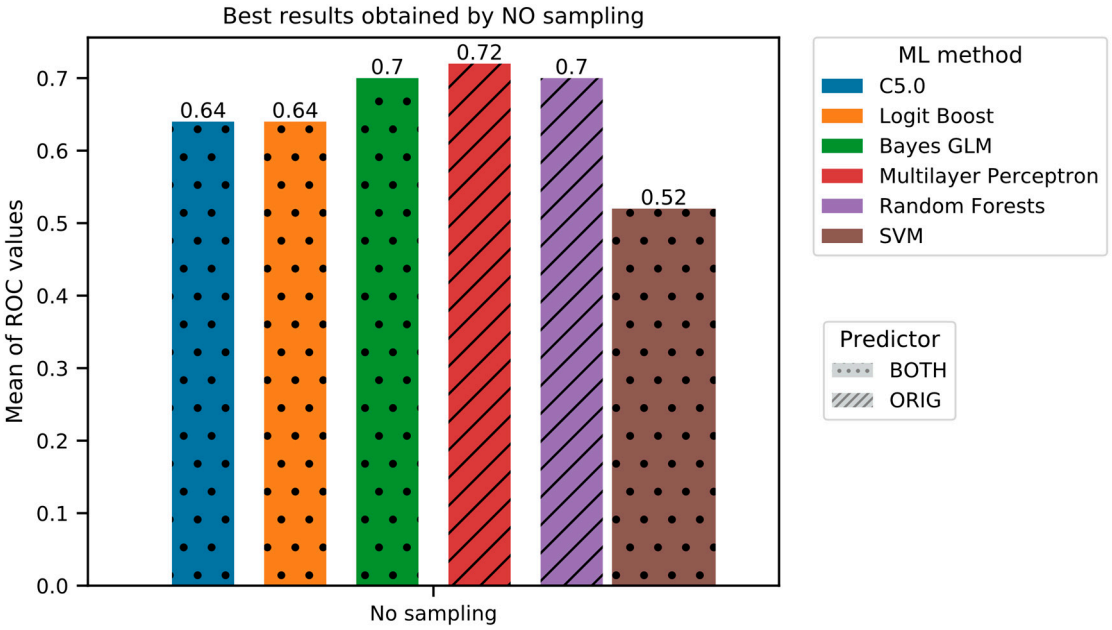

Figure 1. Highest mean accuracy values in non-sampling subset, corresponding to each ML method. Each color bar represents the different classification models. The texture of the bar represents the input predictor from which each model has been generated. The width of the bar represents the number of evaluators that annotated the dataset (3 for the thinner bars and 5 for the thicker one).

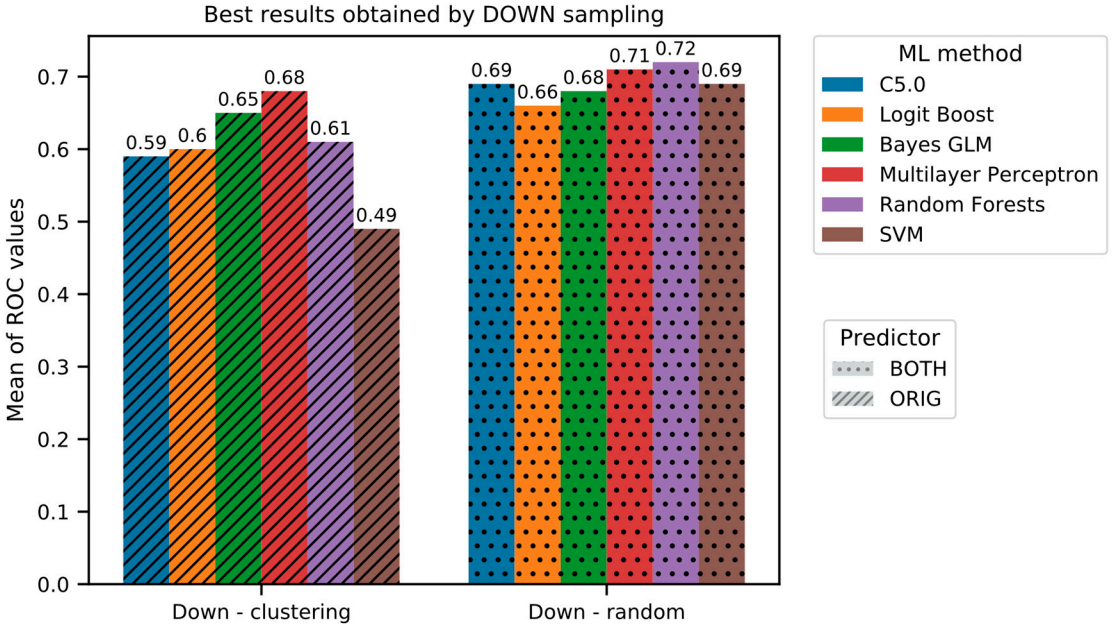

Figure 2. Highest mean accuracy values in down-sampling subset, corresponding to each ML method. Each color bar represents the different classification models. The texture of the bar represents the input predictor from which each model has been generated. The two different down-sampling methods are represented grouped in the X axis.

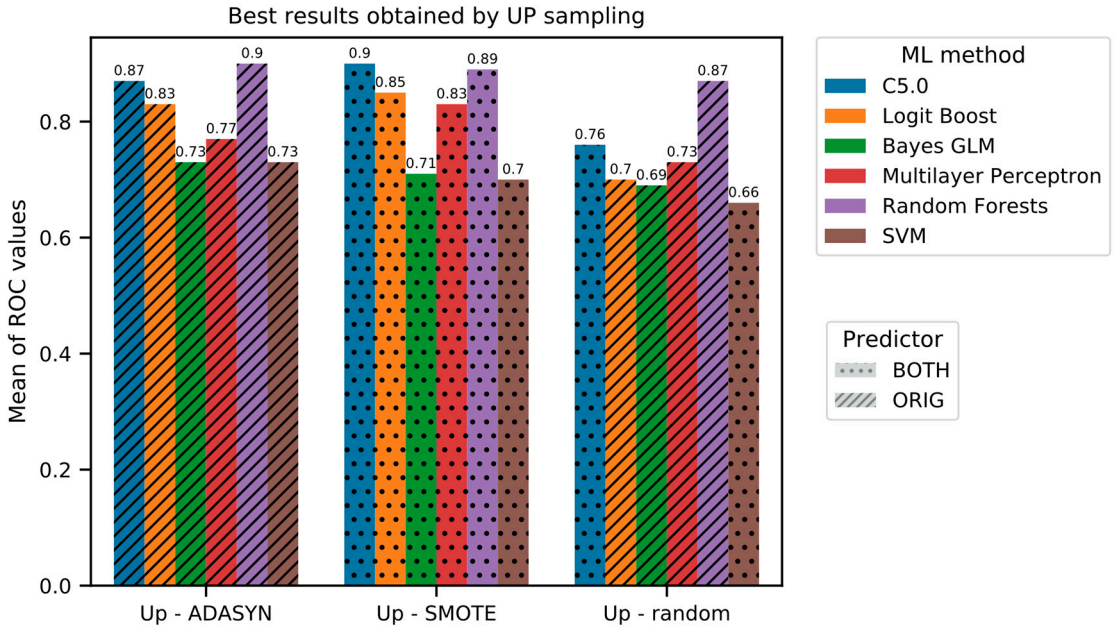

Figure 3. Highest mean accuracy values in up-sampling subset, corresponding to each ML method. Each color bar represents the different classification models. The texture of the bar represents the input predictor from which each model has been generated. The three different up-sampling methods are represented grouped in the X axis.

Table 1. Mean accuracy values for the different sampling subsets, corresponding to each ML method. ML tools are shown from left to right according to higher values of the global average accuracy.

	Random Forests	Multilayer Perceptron	C5.0	Logit Boost	BayesGLM	SVM
No sampling	0.7	0.72	0.64	0.64	0.7	0.52
Down-sampling	0.665	0.695	0.64	0.63	0.665	0.59
Up-sampling	0.89	0.78	0.84	0.79	0.71	0.7
Global average	**0.75**	**0.73**	**0.71**	**0.69**	**0.69**	**0.6**

4. Discussion

The model that provided the highest accuracy from all the studied possibilities was the one generated by the subset obtained from up-sampling with the ADASYN method and corresponding to the Random Forest technique. Such a data subset was formed by the original values of the commercial tools and was annotated by three evaluators.

When no sampling was performed, the accuracy highest values ranged from 0.52 to 0.72. Such accuracies were not much improved when under-sampling (0.49 to 0.72). There was a trend in the accuracy to present larger values when the dataset was balanced via over-sampling the minority class (0.66 to 0.9). The three highest values of the accuracy were 0.9, 0.87 and 0.87, obtained by the three up-sampling methods (Random Forests with ADASYN and C5.0 with SMOTE obtained a mean of ROC values of 0.9; Random Forests with SMOTE got 0.89; and C5.0 with ADASYN got 0.87).

Most of the best results were performed when the dataset was annotated manually by the three evaluators. When the annotation was carried out by the five experts the accuracy mainly decreased. There is just one case when the five evaluators' annotations worked better (SVM without sampling) and five cases when the accuracy was equal to the one obtained by the three evaluators' annotations. Nevertheless, still in those situations, the accuracy was not very high (0.49, 0.52, 0.60, 0.64, 0.66 and 0.76). This can be explained by the disagreement of the three experts versus the five. When considering the five experts, as there were two of them who did not agree with the other three, the dataset increased

its noise levels, and therefore, the accuracy would tend to be lower. On the other hand, adapted values as input to generate models, did not perform as well as the original or the combined ones.

Random Forests was one of the models that obtained better accuracy values for the different sampled subsets (0.7 when no sampling, 0.61 for down-clustering, 0.72 for down-random, 0.9 for ADASYN, 0.89 for SMOTE and 0.87 for up-random). On the contrary, SVM usually got the lower values (0.52 when no sampling, 0.49 for down-clustering, 0.69 for down-random, 0.73 for ADASYN, 0.7 for SMOTE and 0.66 for up-random). The Multilayer Perceptron and C5.0 also obtained high accuracies (the first one when no sampling and for down-clustering, and the second one for SMOTE).

There was a trend in the global accuracy average from the different ML methods: while Random Forests performed the best when up-sampling and regarding the overall accuracy, the Multilayer Perceptron showed the highest values of ROC means when no sampling and down-sampling. Analogously, whilst LogitBoost performed better than BayesGLM when up-sampling, BayesGLM obtained higher results in conditions of no sampling and down-sampling. The same global accuracy average was derived in both cases.

5. Conclusions

The current manuscript presents the research on the different options when analyzing the polarity of tweets regarding a specific set of vaccines and their related diseases. The polarity of tweets published in Twitter in Spanish was annotated by commercial tools and experts. Generally, opinion about vaccines expressed in that social network tend to be non-negative. Therefore, imbalance between classes must be overcome through sampling. Different combinations of sampled classified tweets were used as inputs to generate classification models.

The results showed that the highest accuracy was obtained with the Random Forest model when up-sampling with ADASYN. Over-sampling methods exhibited a better accuracy in most of the cases. However, it must be noted that the analyzed techniques have shown, in some cases, very close results to both the random approaches and other more complex techniques. Those results shown that although the use of more complex techniques such as ADASYN or SMOTE would help to obtain more accurate and stable results based on how they worked, the dataset seems to not have sufficient differences in the distribution of its data to allow such techniques to significantly improve these results in comparison with the random selection of the tweets.

Some other limitations might be pointed out, in particular, the fact that to develop the classifiers, expert manual annotations were needed to provide a trustful labelling. Such experts' annotations may not be always accessible in all the fields if generating a text sentiment classifier in a totally different environment. In future works, more effort should be addressed to assess the classification process with non-binary labels, as it could be "neutral", "negative" and "positive", for instance.

However, the results provided in this paper are significant and allowed us to demonstrate that it is possible to apply sentiment identification techniques by using a metamodel based on different commercial tools even when the number of available tweets for training it is low and unbalanced.

In a more general context, the present work corroborates the fact already stated in literature that over-sampling produces better results than under-sampling. Moreover, this study taught us that there are some ML methods to generate classifiers that outperform others, with Random Forests and the Multilayer Perceptron being the ones with the highest accuracies in comparison to the other ones used. The generated classifiers provide a reasonable solution for annotating tweets in the scope of the mentioned vaccines and related diseases, avoiding the efforts that can require deciding the classification of each tweet manually. Other research questions regarding sentiment classification of texts, wherever they come from, can be solved by following the described or a similar workflow.

Supplementary Materials: The code developed for the current analysis is fully available and accessible at the public repository https://medal.ctb.upm.es/internal/gitlab/mavis/mavis.

Author Contributions: D.F.P.-B. was the main developer of the procedures to extract the knowledge from Twitter. J.M.T. and L.P.S. perform the analysis over the commercial tools and created and evaluated the metamodel created

with the support and guide in the methodology from A.J., M.C., A.J.C.F., A.A., E.M.R., A.G., L.P.S. and A.R.-G. wrote the manuscript with the supervision and revision of all the co-authors. A.R.-G. supervised the whole study and the whole analysis. All authors were involved in the development of this manuscript and gave final approval before submission. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by MSD, Spain under MAVIS Study (VEAP ID: 7789).

Acknowledgments: We would like to thank to the experts that contribute helping in the manual annotation of the tweets for the generation of the machine learning model.

Conflicts of Interest: A.J. and M.C. are full time employees of MSD, sponsor of this study and manufacturer of some of the vaccines subject of this research.

References

1. Diaz, J.A.; Griffith, R.A.; Ng, J.J.; Reinert, S.E.; Friedmann, P.D.; Moulton, A.W. Patients' Use of the Internet for Medical Information. *J. Gen. Intern. Med.* **2002**, *17*, 180–185. [CrossRef] [PubMed]
2. Eysenbach, G. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *J. Med. Internet Res.* **2009**, *11*. [CrossRef]
3. Eysenbach, G.; Powell, J.; Englesakis, M.; Rizo, C.; Steam, A. Health related virtual communities and electronic support groups: Systematic review of the effects of online peer to peer interactions. *BMJ* **2004**, *328*, 1166. [CrossRef]
4. Dugas, A.F.; Hsieh, Y.H.; Levin, S.R.; Pines, J.M.; Mareiniss, D.P.; Mohareb, A.; Gaydos, C.A.; Perl, T.M.; Rothman, R.E. Google flu trends: Correlation with emergency department influenza rates and crowding metrics. *Clin. Infect. Dis.* **2012**, *54*, 463–469. [CrossRef]
5. Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S.; Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **2009**, *457*, 1012–1014. [CrossRef]
6. Gu, Y.; Chen, F.; Liu, T.; Lv, X.; Shao, Z.; Lin, H.; Liang, C.; Zeng, W.; Xiao, J.; Zhang, Y.; et al. Early detection of an epidemic erythromelalgia outbreak using Baidu search data. *Sci. Rep.* **2015**, *5*, 12649. [CrossRef]
7. Wilson, K.; Brownstein, J.S. Early detection of disease outbreaks using the Internet. *CMAJ Can. Med. Assoc. J.* **2009**, *180*, 829–831. [CrossRef]
8. Heymann, D.L.; Rodier, G.R.; WHO operational support team to the global outbreak alert and response network. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. *Lancet Infect. Dis.* **2001**, *1*, 345–353. [CrossRef]
9. Kang, M.; Zhong, H.; He, J.; Rutherford, S.; Yang, F. Using google trends for influenza surveillance in south China. *PLoS ONE* **2013**, *8*, e55205. [CrossRef]
10. M'ikanatha, N.M.; Rohn, D.D.; Robertson, C.; Tan, C.G.; Holmes, J.H.; Kunselman, A.R.; Polachek, C.; Lautenbach, E. Use of the internet to enhance infectious disease surveillance and outbreak investigation. *Biosecur. Bioterror. Biodef. Strategy Pract. Sci.* **2006**, *4*, 293–300. [CrossRef]
11. Polgreen, P.M.; Chen, Y.; Pennock, D.M.; Nelson, F.M. Using internet searches for influenza surveillance. *Clin. Infect. Dis.* **2008**, *47*, 1443–1448. [CrossRef] [PubMed]
12. Collier, N. Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Glob. Public Health* **2012**, *7*, 731–749. [CrossRef] [PubMed]
13. Farooq, A.; Laato, S.; Islam, A.N. impact of online information on self-isolation intention during the COVID-19 pandemic: Cross-sectional study. *J. Med. Internet Res.* **2020**, *22*, e19128. [CrossRef] [PubMed]
14. Ting, D.S.W.; Carin, L.; Dzau, V.; Wong, T.Y. Digital technology and COVID-19. *Nat. Med.* **2020**, *26*, 459–461. [CrossRef]
15. Li, C.; Chen, L.J.; Chen, X.; Zhang, M.; Pang, C.P.; Chen, H. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Eurosurveillance* **2020**, *25*, 2000199. [CrossRef]
16. Bossetta, M. The digital architectures of social media: Comparing political campaigning on facebook, twitter, instagram, and snapchat in the 2016 U.S. election. *J. Mass Commun. Q.* **2018**, *95*, 471–496. [CrossRef]
17. Bello, G.; Menéndez, H.; Okazaki, S.; Camacho, D. Extracting collective trends from twitter using social-based data mining. In *Computational Collective Intelligence. Technologies and Applications*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 622–630. [CrossRef]

18. Wang, C.; Chen, W.; Wang, Y. Scalable influence maximization for independent cascade model in large-scale social networks. *Data Min. Knowl. Discov.* **2012**, *25*, 545–576. [CrossRef]
19. Das, K.; Acharjya, D.P.; Patra, M.R. Opinion mining about a product by analyzing public tweets in twitter. In Proceedings of the 2014 International Conference on Computer Communication and Informatics, Coimbatore, India, 3–5 January 2014; pp. 1–4. [CrossRef]
20. Chamlertwat, W.; Bhattarakosol, P.; Rungkasiri, T.; Haruechaiyasak, C. Discovering consumer insight from twitter via sentiment analysis. *J. UCS* **2012**. [CrossRef]
21. Asur., S.; Huberman, B.A. Predicting the future with social media. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, ON, Canada, 31 August–3 September 2010; pp. 492–499. [CrossRef]
22. Souza, T.T.P.; Kolchyna, O.; Treleaven, P.C.; Aste, T. Twitter sentiment analysis applied to finance: A case study in the retail industry. *arXiv* **2015**, arXiv:1507.00784. Available online: http://arxiv.org/abs/1507.00784 (accessed on 3 June 2020).
23. Yang, S.Y.; Mo, S.Y.K.; Liu, A. Twitter financial community sentiment and its predictive relationship to stock market movement. *Quant. Finance* **2015**, *15*, 1637–1656. [CrossRef]
24. Chunara, R.; Andrews, J.R.; Brownstein, J.S. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *Am. J. Trop. Med. Hyg.* **2012**, *86*, 39–45. [CrossRef] [PubMed]
25. Culotta, A. Towards detecting influenza epidemics by analyzing twitter messages. In Proceedings of the First Workshop on Social Media Analytics, New York, NY, USA, 25 July 2010; pp. 115–122. [CrossRef]
26. Chew, C.; Eysenbach, G. Pandemics in the age of twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS ONE* **2010**, *5*, e14118. [CrossRef] [PubMed]
27. Freifeld, C.C.; Brownstein, J.S.; Menone, C.M.; Bao, W.; Filice, R.; Kass-Hout, T.; Dasgupta, N. Digital drug safety surveillance: Monitoring pharmaceutical products in twitter. *Drug Saf.* **2014**, *37*, 343–350. [CrossRef] [PubMed]
28. Curtis, J.R.; Chen, L.; Higginbotham, P.; Nowell, W.B.; Gal-Levy, R.; Willig, J.; Safford, M.; Coe, J.; O'Hara, K.; Sa'adon, R. Social media for arthritis-related comparative effectiveness and safety research and the impact of direct-to-consumer advertising. *Arthritis Res. Ther.* **2017**, *19*, 48. [CrossRef]
29. Zhou, X.; Coiera, E.; Tsafnat, G.; Arachi, D.; Ong, M.-S.; Dunn, A.G. Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter. *Stud. Health Technol. Inform.* **2015**, *216*, 761–765.
30. Salathé, M.; Khandelwal, S. Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Comput. Biol.* **2011**, *7*, e1002199. [CrossRef]
31. Broniatowski, D.A.; Jamison, A.M.; Qi, S.; AlKulaib, L.; Chen, T.; Benton, A.; Quinn, S.C.; Dredze, M. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *Am. J. Public Health* **2018**, *108*, 1378–1384. [CrossRef]
32. Kata, A. A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine* **2010**, *28*, 1709–1716. [CrossRef]
33. Tomeny, T.S.; Vargo, C.J.; El-Toukhy, S. Geographic and demographic correlates of autism-related anti-vaccine beliefs on Twitter, 2009–2015. *Soc. Sci. Med.* **1982**, *191*, 168–175. [CrossRef]
34. Becker, F.H.; Larson, H.J.; Bonhoeffer, J.; van Mulligen, E.M.; Kors, J.A.; Sturkenboom, M.C.J.M. Evaluation of a multinational, multilingual vaccine debate on Twitter. *Vaccine* **2016**, *34*, 6166–6171. [CrossRef]
35. Dunn, G.; Leask, J.; Zhou, X.; Mandl, K.D.; Coiera, E. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: An observational study. *J. Med. Internet Res.* **2015**, *17*, e144. [CrossRef] [PubMed]
36. Luo, X.; Zimet, G.; Shah, S. A natural language processing framework to analyse the opinions on HPV vaccination reflected in twitter over 10 years (2008–2017). *Hum. Vaccines Immunother.* **2019**, *15*, 1496–1504. [CrossRef] [PubMed]
37. Massey, P.M.; Leader, A.; Yom-Tov, E.; Budenz, A.; Fisher, K.; Klassen, A.C. Applying multiple data collection tools to quantify human papillomavirus vaccine communication on twitter. *J. Med. Internet Res.* **2016**, *18*, e318. [CrossRef] [PubMed]

38. Shapiro, G.K.; Surian, D.; Dunn, A.G.; Perry, R.; Kelaher, M. Comparing human papillomavirus vaccine concerns on Twitter: A cross-sectional study of users in Australia, Canada and the UK. *BMJ Open* **2017**, *7*, e016869. [CrossRef]
39. Du, J.; Xu, J.; Song, H.-Y.; Tao, C. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with twitter data. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 69. [CrossRef]
40. Keim-Malpass, J.; Mitchell, E.M.; Sun, E.; Kennedy, C. Using twitter to understand public perceptions regarding the #HPV vaccine: Opportunities for public health nurses to engage in social marketing. *Public Health Nurs.* **2017**, *34*, 316–323. [CrossRef]
41. Amith, M.; Cohen, T.; Cunningham, R.; Savas, L.S.; Smith, N.; Cuccaro, P.; Gabay, E.; Boom, J.; Schvaneveldt, R.; Tao, C. Mining HPV vaccine knowledge structures of young adults from reddit using distributional semantics and pathfinder networks. *Cancer Control J. Moffitt Cancer Cent.* **2020**, *27*. [CrossRef]
42. Suppli, H.; Hansen, N.D.; Rasmussen, M.; Valentiner-Branth, P.; Krause, T.G.; Mølbak, K. Decline in HPV-vaccination uptake in Denmark—The association between HPV-related media coverage and HPV-vaccination. *BMC Public Health* **2018**, *18*, 1360. [CrossRef]
43. Ortiz, R.R.; Smith, A.; Coyne-Beasley, T. A systematic literature review to examine the potential for social media to impact HPV vaccine uptake and awareness, knowledge, and attitudes about HPV and HPV vaccination. *Hum. Vaccines Immunother.* **2019**, *15*, 1465–1475. [CrossRef]
44. Aramaki, E.; Maskawa, S.M.; Morita, M. Twitter catches the flu: Detecting influenza epidemics using Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 16–20 November 2011; pp. 1568–1576. Available online: https://dl.acm.org/doi/abs/10.5555/2145432.2145600 (accessed on 2 June 2020).
45. Signorini, A.; Segre, A.M.; Polgreen, P.M. The use of twitter to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 pandemic. *PLoS ONE* **2011**, *6*, e19467. [CrossRef]
46. Wakamiya, S.; Kawai, Y.; Aramaki, E. Twitter-based influenza detection after flu peak via tweets with indirect information: Text mining study. *JMIR Public Health Surveill.* **2018**, *4*, e65. [CrossRef] [PubMed]
47. Sharma, M.; Yadav, K.; Yadav, N.; Ferdinand, K.C. Zika virus pandemic-analysis of Facebook as a social media health information platform. *Am. J. Infect. Control* **2017**, *45*, 301–302. [CrossRef] [PubMed]
48. Ghenai, A.; Mejova, Y. Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter. In Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, USA, 23–26 August 2017; p. 518. [CrossRef]
49. Christakis, N.A.; Fowler, J.H. The spread of obesity in a large social network over 32 Years. *N. Engl. J. Med.* **2007**, *357*, 370–379. [CrossRef] [PubMed]
50. Zhang, Y.; He, D.; Sang, Y. Facebook as a platform for health information and communication: A case study of a diabetes group. *J. Med. Syst.* **2013**, *37*, 9942. [CrossRef] [PubMed]
51. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
52. Nakov, P.; Ritter, A.; Rosenthal, S.; Sebastiani, F.; Stoyanov, V. SemEval-2016 task 4: Sentiment analysis in twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16–17 June 2016; pp. 1–18. [CrossRef]
53. Smailović, J.; Grčar, M.; Lavrač, N.; Žnidaršič, M. Stream-based active learning for sentiment analysis in the financial domain. *Inf. Sci.* **2014**, *285*, 181–203. [CrossRef]
54. González, A.R.; Tuñas, J.M.; Peces-Barba, D.F.; Ruiz, E.M.; Jaramillo, A.; Cotarelo, M.; Conejo, A.; Arce, A.; Gil, A. Creating a metamodel based on machine learning to identify the sentiment of vaccine and disease-related messages in Twitter: The MAVIS study. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MI, USA, 28–30 July 2020; p. 6.
55. Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C. New avenues in opinion mining and sentiment analysis. *IEEE Intell. Syst.* **2013**, *28*, 15–21. [CrossRef]
56. Nasukawa, T.; Yi, J. Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture, Sanibel Island, FL, USA, 23–25 October 2003; pp. 70–77. [CrossRef]
57. Branco, P.; Torgo, L.; Ribeiro, R.P. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.* **2016**, *49*. [CrossRef]

58. Krawczyk, B.; McInnes, B.T.; Cano, A. Sentiment classification from multi-class imbalanced twitter data using binarization. In Proceedings of the International Conference on Hybrid Artificial Intelligence Systems, Cham, Germany, 2 June 2017; pp. 26–37. [CrossRef]
59. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
60. Maheshwari, S.; Jain, D.R.C.; Jadon, D.R.S. A review on class imbalance problem: Analysis and potential solutions. *Int. J. Comput. Sci. Issues (IJCSI)* **2017**. [CrossRef]
61. Drummond, C.; Holte, R.C. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling. In *Workshop on Learning from Imbalanced Datasets II*; Citeseer: Washington, DC, USA, 2003; pp. 1–8.
62. Yen, S.-J.; Lee, Y.S. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst. Appl.* **2009**, *36*, 5718–5727. [CrossRef]
63. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
64. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328. [CrossRef]
65. Information on See5/C5.0. Available online: https://www.rulequest.com/see5-info.html (accessed on 17 June 2020).
66. Quinlan, J.R. *Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.
67. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* **2000**, *28*, 337–407. [CrossRef]
68. Dettling, M.; Bühlmann, P. Boosting for tumor classification with gene expression data. *Bioinformatics* **2003**, *19*, 1061–1069. [CrossRef]
69. Gelman, A.; Jakulin, A.; Pittau, M.G.; Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2008**, *2*, 1360–1383. [CrossRef]
70. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408. [CrossRef]
71. Hastie, T.; Tibshirani, R.; Friedman, J. Neural networks. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Hastie, T., Tibshirani, R., Friedman, J., Eds.; Springer: New York, NY, USA, 2009; pp. 389–416.
72. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
73. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
74. Carnevale, L.; Celesti, A.; Fiumara, G.; Galletta, A.; Villari, M. Investigating classification supervised learning approaches for the identification of critical patients' posts in a healthcare social network. *Appl. Soft Comput.* **2020**, *90*, 106155. [CrossRef]