Some Like it Hoax: Automated Fake News Detection in Social Networks

Eugenio Tacchini[1], Gabriele Ballarin[2], Marco L. Della Vedova[3], Stefano Moret[4], and Luca de Alfaro[5]

[1] Università Cattolica, Piacenza, Italy eugenio.tacchini@unicatt.it
[2] Independent researcher gabriele.ballarin@gmail.com
[3] Università Cattolica, Brescia, Italy marco.dellavedova@unicatt.it
[4] École Polytechnique Fédérale de Lausanne, Switzerland moret.stefano@gmail.com
[5] Department of Computer Science, UC Santa Cruz, CA, USA. luca@ucsc.edu

Abstract. In the recent years, the reliability of information on the Internet has emerged as a crucial issue of modern society. Social network sites (SNSs) have revolutionized the way in which information is spread by allowing users to freely share content. As a consequence, SNSs are also increasingly used as vectors for the diffusion of misinformation and hoaxes. The amount of disseminated information and the rapidity of its diffusion make it practically impossible to assess reliability in a timely manner, highlighting the need for automatic online hoax detection systems.

As a contribution towards this objective, we show that Facebook posts can be classified with high accuracy as hoaxes or non-hoaxes on the basis of the users who "liked" them. We present two classification techniques, one based on logistic regression, the other on a novel adaptation of boolean crowdsourcing algorithms. On a dataset consisting of 15,500 Facebook posts and 909,236 users, we obtain classification accuracies exceeding 99% even when the training set contains less than 1% of the posts. We further show that our techniques are robust: they work even when we restrict our attention to the users who like both hoax and non-hoax posts. These results suggest that mapping the diffusion pattern of information can be a useful component of automatic hoax detection systems.

1 Introduction

The World Wide Web (WWW) has revolutionized the way in which information is disseminated. In particular, social network sites (SNSs) are platforms where content can be freely shared, enabling users to actively participate to - and, possibly, influence - information diffusion processes. As a consequence, SNSs are also increasingly used as vectors for the dissemination of spam [1], conspiracy theories and *hoaxes*, i.e. intentionally

crafted fake information. This recently led to the emphatic definition of our current times as the *age of misinformation* [2]. A significant share of hoaxes on SNSs diffuses rapidly, with a peak in the first 2 hours [3]. This finding, together with the high amount of shared content, highlights the need of automatic online hoax detection systems [4].

In the literature, various approaches have been proposed for automatic hoax detection, covering quite heterogeneous applications. Historically, one of the first applications has been hoax detection in e-mail messages and webpages. In the context of scam e-mail detection, spamassassin uses keyword-based methods with logistic regression [5]; Petković et al. [6] and Ishak et al. [7] proposed the use of distance-based methods; Vuković et al. [8] applied neural network and advanced text processing; Yevseyeva et al. [9] used evolutionary algorithms for the development of anti-spam filters. Sharifi et al. [10] applied logistic regression to automatically detect scam on webpages, reaching an accuracy of 98%.

The concepts of trust and reputation [11, 12] can be adopted for hoax detection in applications with a dominant social component. Metrics and algorithms for this purpose have been proposed by Golbeck and Hendler [13]. Adler and de Alfaro [14] developed a content-driven user reputation system for Wikipedia, allowing to predict the quality of new contributions. The detection of Wikipedia hoaxes has been addressed e.g. in [15, 16, 17]. More recently, automatic hoax detection in SNSs has gained increasing interest. As an example, Chen et al. [18] developed a semi-supervised scam detector for Twitter based on self-learning and clustering analysis, while Ito et al. [19] proposed the use of Latent Dirichlet Allocation (LDA) to assess the credibility of tweets.

The key idea behind our work, which constitutes its main novelty, is that hoaxes can be identified with great accuracy on the basis of the users that interact with them. In particular, focusing on Facebook, we answer the following research question: *Can a hoax be identified based on the users who "liked" it?* We consider a dataset consisting of 15,500 posts and 909,236 users; the posts originate from pages that deal with either scientific topics or with conspiracies and fake scientific news [2]. We propose two classification techniques. One consists in applying logistic regression, considering the user interaction with posts as features. The other technique consists in a novel adaptation of boolean label crowdsourcing techniques to a setting where a training set is available, but no prior assumption on users being mostly reliable can be made.

The proposed techniques yield an accuracy exceeding 99% even for training sets consisting of less of 1% of posts. These results are obtained

in spite of the fact that the communities of users participating in the scientific and conspiracy pages overlap. Our main contributions, in summary, are: *i*) the proposal of a novel way to identify hoaxes on SNSs based on the users who interacted with them rather than their content; *ii*) an improved version of the harmonic crowdsourcing method, suited to hoax detection in SNSs; *iii*) the application on Facebook and, in particular, on a representative dataset obtained from the literature.

The code we developed for this paper is available from `https://github.com/gabll/some-like-it-hoax`.

2 Dataset

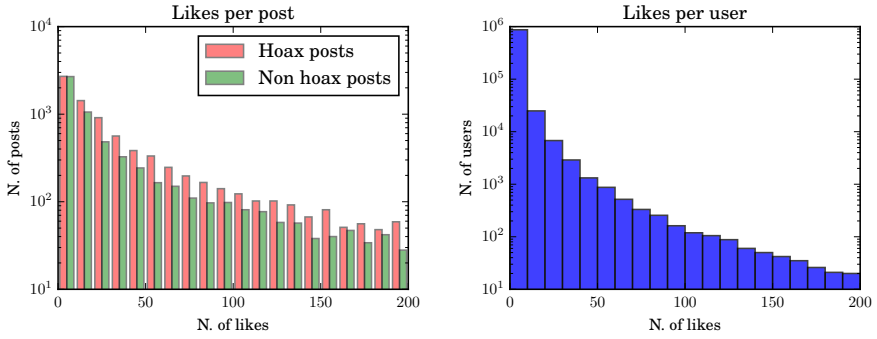
Our dataset consists in all the public posts and posts' likes of a list of selected Facebook pages during the second semester of 2016: from Jul. 1st, 2016 to Dec. 31st, 2016. We collected the data by means of the Facebook Graph API[6] on Jan. 27th, 2017.

We based our selection of pages on [2]. In that work, the authors present a list of Facebook pages divided into two categories: scientific news sources vs. conspiracy news sources. We assume all posts from scientific pages to be reliable, i.e. "non-hoaxes", and all posts from conspiracy pages to be "hoaxes". Among the 73 pages listed in [2], we limited our analysis to the top 20 pages of both categories. It is worth noting that at the time of data collection, not all the pages were still available: some of them had been deleted in the meantime, or were no longer publicly accessible. We note also that the actual posts comprising our dataset are distinct from those originally included in the dataset of [2], as we performed our data collection in a different, and more recent, period.

The resulting dataset, the so-called *complete dataset*, is composed of 15,500 posts from 32 pages (14 conspiracy and 18 scientific), with more than 2,300,00 likes by 900,000+ users (Table 1). Among posts, 8,923 (57.6%) are hoaxes and 6,577 (42.4%) non-hoaxes.

As a first observation, the distribution of the number of likes per post is exponential-like, as attested by the histograms in Fig. 1 (a); the majority of the posts have few likes. Hoax posts have, on average, more likes than non-hoax posts. In particular, some figures about the number of likes per post are: average, 204.5 (for hoax post) vs. 84.0 (non-hoax); median, 22 (hoax) vs. 14 (non-hoax); maximum, 121,491 (hoax) vs. 13,608 (non-hoax).

[6] See `https://developers.facebook.com/docs/graph-api`. We used version 2.6.



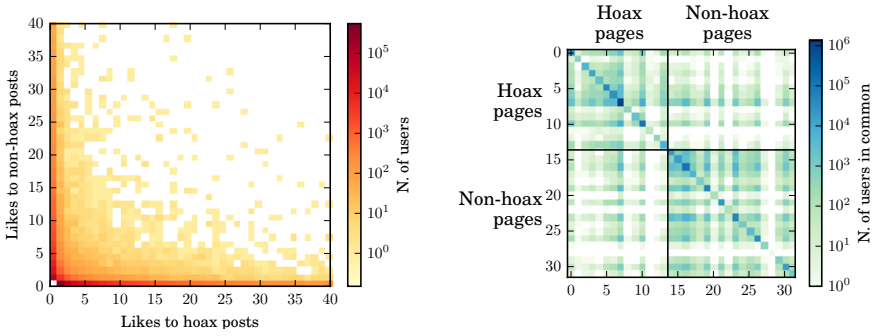
(a) (b)

Fig. 1: Likes per post (a) and likes per user (b) histograms for the dataset. Plots are represented with semi-logarithmic scale and are tuncated to 200 likes.

	N. posts	N. users	N. likes
Complete	15,500	909,236	2,376,776
Intersection	10,520	14,139	117,641

Table 1: Composition of complete and intersection datasets.

A second observation is related to the number of likes per user: once again, Fig. 1 (b) shows an exponential-like distribution. The majority of the users appears in the dataset with one single like (629,146 users, 69.2%), while the maximum number of likes by a user is 1,028. Users can be divided into three categories based on what they liked: *i*) those who liked hoax posts only, *ii*) those who liked non-hoax posts only, and *iii*) those who liked at least one post belonging to a hoax page, and one belonging to a non-hoax page. Fig. 2 (a) shows that, despite a high polarization, there are many users in the mixed category: among users with at least 2 likes, 209,280 (74.7%) liked hoax post only, 56,671 (20.3%) liked non-hoax post only, and 14,139 (5.0%) are in the mixed category. This latter category gives rise to the *intersection dataset*, which consists only of the users who liked *both* hoax and non-hoax posts, and of the posts these users liked. The intersection dataset was introduced to study the performance of our methods for communities of users that are not strongly polarized towards hoax or non-hoax posts, as will be discussed in Section 4. The composition of the intersection dataset is summarized in Table 1.



(a) (b)

Fig. 2: Users characterization: hoax vs. non-hoax likes per user heat-map (a) and users in common between pages (b)

A third observation concerns the relation between pages, measured by the number of users that pages have in common: given each pair of pages, we study how many users liked at least one post from one page and one post from the other page. Fig. 2 (b) shows the result as a symmetric matrix: each page vs. each other page. Color intensity displays that hoax pages have more users in common with other hoax pages (up-left part, which appears darker) than with non-hoax pages (up-right and bottom-left). The same applies to non-hoax pages (bottom-right). Nevertheless, the figure shows that the communities gravitating around hoax and non-hoax pages share many common users (as evidenced also from the composition of the intersection dataset).

3 Algorithmic Classification of Posts

Our goal is to classify posts into *hoax* and *non-hoax* posts. According to the analysis of social media sharing by [3], "users tend to aggregate in communities of interest, which causes reinforcement and fosters confirmation bias, segregation, and polarization", and "users mostly tend to select and share content according to a specific narrative and to ignore the rest." This suggests that the set of users who like a post should be highly indicative of the nature of the post. We present two approaches, one based on logistic regression, the other based on boolean crowdsourcing algorithms.

3.1 Classification via logistic regression

We formulate the post classification problem as a supervised learning, binary classification problem. We consider a set of posts I and a set of users U. Each post $i \in I$ has an associated set of features $\{x_{iu} \mid u \in U\}$, where $x_{iu} = 1$ if u liked post i, and $x_{iu} = 0$ otherwise. We classify the posts on the basis of their features, that is, on the basis of which users liked them.

To perform the classification, we use a logistic regression model. The logistic regression model learns a weight w_u for each user $u \in U$; the probability p_i that a post i is non-hoax is then given by $p_i = 1/(1+e^{-y_i})$, where $y_i = \sum_{u \in U} x_{iu} w_u$. Intuitively, $w_u > 0$ (resp. $w_u < 0$) indicates that u likes mostly non-hoax (resp. hoax) posts.

We chose logistic regression for two reasons. First, logistic regression is well suited to problems with a very large, and uniform, set of features. In our case, we have about a million features (users) in our dataset, but a real application would involve up to hundreds of millions of users. Second, our logistic regression setting enjoys a *non-interference* property with respect to unrelated set of users that facilitates learning, and is appealing on conceptual grounds. Specifically, assume that the set of users and posts are partitioned into disjoint subsets $U = U_1 \cup U_2$, $I = I_1 \cup I_2$, so that users in U_k like only posts in I_k, for $k = 1, 2$. This situation can arise, for instance, when there are two populations of users and posts in different languages, or simply when two topics are very unrelated. In such a setting, it is equivalent to train a single model, or to train separately two models, one for I_1, U_1, one for I_2, U_2, and then take their "union". This because the weights w_u for $u \in U_{3-k}$ do not matter for classifying posts in I_k, $k = 1, 2$, since the features x_{iu} with $i \in I_k$ and $u \in U_{3-k}$ are all zero. In other words, models for unrelated communities do not interfere: if we learn a model for I_1, U_1, we do not need to revise the model once the community I_2, U_2 is discovered: all we need to do is learn a model of this second community, and use it jointly with the first.

3.2 Classification via harmonic boolean label crowdsourcing

The weak aspect of logistic regression is that it does not transfer information across users who liked some of the same posts. In particular, if the training set does not contain any post liked by a user u, then logistic regression will not be able to learn anything about u, and w_u will be undetermined. Thus, posts that are only liked by users not in the training set cannot be classified. As an alternative approach, we propose to

perform the hoax/non-hoax classification using algorithms derived from crowdsourcing, and precisely, from the *boolean label crowdsourcing* (BLC) problem.

In the BLC problem, users provide True/False labels for posts, indicating for instance whether a post is vandalism, or whether it violates community guidelines. The BLC problem consists in computing the consensus labels from the user input [20, 21, 22]. We model liking a post as voting True on that post.

Our setting differs from standard BLC in one important respect. Standard BLC algorithms do not use a learning set: rather, they assume that people are more likely to tell the truth than to lie. The algorithms compare what people say, correct for the effect of the liars, and reconstruct a consensus truth [20, 22]. In our setting, we cannot assume that users are more likely to tell the truth, that is, like preferentially non-hoax posts. Indeed, hoax articles may well have more "likes" than non-hoax ones. Rather, we will rely on a learning set of posts for which the ground truth is known.

We present here an adaptation of the *harmonic* algorithm of [22] to a setting with a learning set of posts. We chose the harmonic algorithm because it is computationally efficient, can cope with large datasets, and it offers good accuracy in practice, as evidenced in [22]. Furthermore, while the harmonic algorithm can be adapted to the presence of a learning set, it is less obvious how to do so for some of the other algorithms, such as those of [20].

We represent the dataset as a bipartite graph $(I \cup U, L)$, where $L \subseteq I \times U$ is the set of likes. We denote by $\partial i = \{u \mid (i,u) \in L\}$ and $\partial u = \{i \mid (i,u) \in L\}$ the 1-neighborhoods of a post $i \in I$ and user $u \in U$, respectively.

The harmonic algorithm maintains for each node $v \in I \cup U$ two non-negative parameters α_v, β_v. These parameters define a beta distribution: intuitively, for a user u, $\alpha_u - 1$ represents the number of times we have seen the user like a non-hoax post, and $\beta_u - 1$ represents the number of times we have seen the user like a hoax post. For a post i, $\alpha_i - 1$ represents the number of non-hoax votes it has received, and $\beta_i - 1$ represents the number of hoax votes it has received. For each node v, let $p_v = \alpha_v/(\alpha_v + \beta_v)$ be the mean of its beta distribution: for a user u, p_u is the (average) probability that the user is truthful (likes non-hoax posts), and for a post i, p_i is the (average) probability that i is not a hoax. Letting $q_v = 2p_v - 1 = (\alpha_v - \beta_v)/(\alpha_v + \beta_v)$, positive values of q_v indicate propensity for non-hoax, and negative values, propensity for hoax.

Let the training set consist of two subets $I_H, I_N \subseteq I$ of known hoax and non-hoax posts. The algorithm sets $q_i := -1$ for all $i \in I_H$, and $q_i := 1$ for all $i \in I_N$; it sets $q_i = 0$ for all other posts $i \in I \setminus (I_H \cup I_N)$. The algorithm then proceeds by iterative updates. First, for each user $u \in U$, it lets:

$$\alpha_u := A + \sum\{q_i \mid i \in \partial u, q_i > 0\} \qquad \beta_u := B - \sum\{q_i \mid i \in \partial u, q_i < 0\}$$
$$q_u := (\alpha_u - \beta_u)/(\alpha_u + \beta_u)\ . \tag{1}$$

The positive constants A, B determine the amount of evidence needed to sway the algorithm towards believing that a user likes hoax or non-hoax posts: the higher the values of A and B, the more evidence will be required. After some experimentation, we settled on the values $A = 5.01$ and $B = 5$, corresponding to a very weak a-priori preference of users for non-hoax posts. This corresponds to needing about 5 "likes" from known good (resp bad) users to reach a 2:1 probability ratio in favor of non-hoax (resp. hoax), which seems intuitively reasonable. The algorithm then updates the values for each post $i \in I \setminus (I_H \cup I_N)$ by:

$$\alpha_i := A' + \sum\{q_u \mid u \in \partial i, q_u > 0\} \qquad \beta_i \ := B' - \sum\{q_u \mid u \in \partial i, q_u < 0\}$$
$$q_i := (\alpha_i - \beta_i)/(\alpha_i + \beta_i)\ . \tag{2}$$

We choose $A' = B' = 5$, thus adopting a symmetrical a-priori for items being hoax vs. non-hoax. The updates (1)–(2) are performed iteratively; while they could be performed until a fixpoint is reached, we just perform them 5 times, as further updates do not yield increased accuracy. Finally, we classify a post i as hoax if $q_i < 0$, and as non-hoax otherwise.

The harmonic algorithm satisfies the *non-interference* property described for logistic regression, since information is only propagated along graph edges that correspond to *"likes"*.

The harmonic algorithm is able to propagate information from posts where the ground truth is known, to posts that are connected by common users. In the first iteration, the users who liked mostly hoax (resp. non-hoax) posts will see their β (resp. α) coefficient increase, and thus their preferences will be characterized. In the next iteration, the user preferences will be reflected on post beliefs, and these post beliefs will subsequently be used to infer the preferences of more users, and so on. We will see how the ability to transfer information will allow the harmonic algorithm to reach high levels of accuracy even starting from small training sets.

4 Results

We characterize the performance of the logistic regression and harmonic BLC algorithm via two sets of experiments. The first set of experiments measures the accuracy of the algorithms as a function of the number of posts available as training set. Since the training set can be produced, in general, only via a laborious process of manual post inspection, these results tell us how much do we need to invest in manual labeling, to reap the benefits of automated classification. The second set of experiments measures how much information our learning is able to transfer from one set of pages to another. As the community of Facebook users is organized around pages, these experiments shed light on how much what we learn from one community can be transferred to another, via the shared users among communities.

4.1 Accuracy of classification vs. training set size

Cross-validation analysis. We performed a standard cross-validation analysis of logistic regression and of the harmonic algorithm for BLC. The cross-validation was performed by dividing the posts in the dataset into 80% training and 20% testing, and performing a 5-fold cross-validation analysis. Both approaches performed remarkably well, with accuracies exceeding 98.6% for logistic regression and 99.4% for the harmonic algorithm.

Accuracy vs. training set size. Cross-validation is not the most insightful evaluation of our algorithms. In classifying news posts as hoax or non-hoax, there is a cost involved in creating the training set, as it may be necessary to examine each post individually. The interesting question is not the level of accuracy we can reach when we know the ground truth for 80% of the posts, but rather, how large a training set do we need in order to reach a certain level of accuracy. In order to be able to scale up to the size of social network information sharing, our approaches need to be able to produce an accurate classification relying on a small fraction of posts of known class.

To better understand this point, it helps to contrast the situation for standard ML settings, versus our post-classification problem. In standard ML settings, the set of features is chosen in advance, and the model that is developed from the 80% of data in the training set is expected to be useful for all future data, and not merely the 20% that constitutes the evaluation set. Thus, cross-validation provides a measure of performance

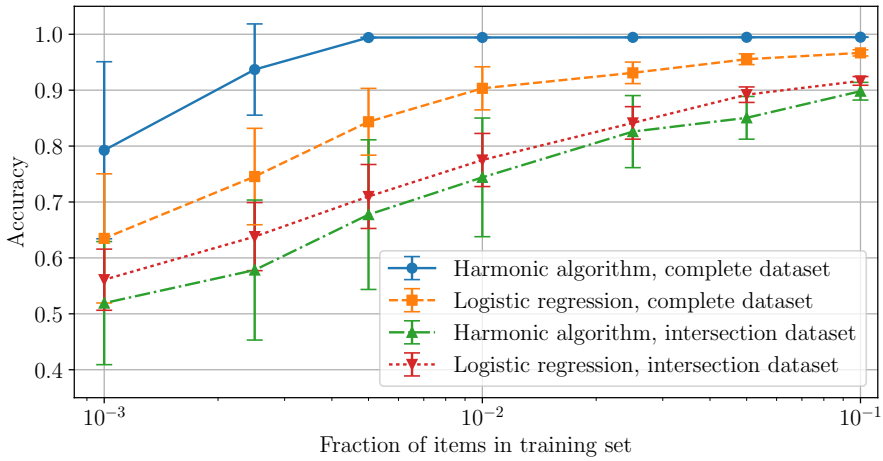


Fig. 3: Accuracy of the logistic and harmonic BLC algorithms on the complete and intersection datasets, as a function of the fraction of posts in the training set. The data is the average of 50 runs; the error bars denote the standard deviation of the accuracy value of each run.

for any future data. In contrast, in our setting the "features" consist in the users that liked the posts. The larger the set of posts we consider, the larger the set of users that might have interacted with them; we cannot assume that the model developed from 80% of our data will be valid for any set of future posts to be classified. Rather, the interesting question is, how many posts do we need to randomly select and classify, in order to be able to automatically classify all others?

We report the classification accuracy both for the complete dataset, and for the intersection dataset. The intersection dataset (defined in Section 2) allows to study the performance of our methods for communities of users that are not strongly polarized towards hoax or non-hoax posts.

In Fig. 3, we report the accuracy our methods as a function of the size of the training set. In the figure, the classification accuracy reported for each training set size is the average of 50 runs. In each run, we select randomly a subset of posts to serve as training set, and we measure the classification accuracy on all other posts. The error bars in the figure denote the standard deviation of the classification accuracy of each run. Thus, the error bars provide an indication of run-to-run variablity (how much the accuracy varies with the particular training set), rather than of

the precision in measuring the average accuracy. The standard deviation with which the average accuracy is known is about seven times smaller.

For the complete dataset, the harmonic BLC algorithm is the superior one. As long as the training set contains at least 0.5% of the posts, or about 80 posts, the accuracy exceeds 99.4%. For even lower training set sizes the accuracy decreases, but it is still about 80% for a training set consisting of 0.1% of posts, or about 15 posts. Logistic regression is somewhat inferior, but still yields accuracy above 90% for training sets consisting of only 1% of the posts.

On the intersection dataset, on the other hand, the logistic regression approach is the superior one. While the differences between the logistic regression and harmonic BLC algorithms is not large, the performance of logistic regression starts at 91.6% for a training set consisting of 10% of posts, and degrades towards 56% for a training set consisting of 0.1% of posts, maintaining a performance margin of 3–4% over harmonic BLC.

Generally, these results indicate that harmonic BLC is more efficient at transfering information across the dataset. Its inferior performance for the intersection dataset may be explained by the fact that the artificial construction of the intersection dataset biases towards the transfer of erroneous information. Most users have only a few likes (see Figure 1). The intersection dataset filters out all users who liked only one post, and of the users who liked two posts, the intersection dataset filters out all those who liked two posts of the same hoax/non-hoax class. As a consequence, the intersection dataset heavily over-samples "straddling" users who like exactly two posts, one hoax, one not; these straddling users constitute 32% of the users in the intersection dataset. When the two posts liked by a straddling user belong one to the training, one to the evaluation dataset, the straddling user contributes in the wrong direction to the classification of the post in the evaluation set.

4.2 Cross-page learning

As the community of Facebook users naturally revolves around common interests and pages, an interesting question concerns whether what we learn from one community of users on one page transfers to other pages. In order to answer this question, we test our classifiers on posts related to pages that they have not seen during the training phase. This further allows to assess the validity of the proposed method in real-world situations, in which the system will need to detect fake news in new pages, i.e. pages not belonging to the ground truth. To this end, we perform two experiments in which the set of pages from which we learn, and those

	One-page-out		Half-pages-out	
	Avg accuracy	Stdev	Avg accuracy	Stdev
Logistic regression	0.794	0.303	0.716	0.143
Harmonic BLC	0.991	0.023	0.993	0.002

Table 2: Accuracy (fraction of correctly classified posts) when leaving one page out, and when leaving out half of the pages, from the training set. For one-page-out, we report average and standard deviation obtained by leaving out each page in turn. For half-pages-out, we report average and standard deviation of the accuracy obtained in 50 runs.

on which we test, are disjoint. In the first experiment, *one-page-out,* we select in turn each page, and we place all its posts in the testing set; the posts belonging to all other pages are in the training set. In the second experiment, *half-pages-out,* we perform 50 runs. In each run, we randomly select a set consisting of half of the pages in the dataset, and we place the posts belonging to those pages in the testing set, and all others in the training set. The results are reported in Table 2.

The results clearly indicate that the harmonic BLC algorithm is the superior one for transferring information across pages, achieving essentially perfect accuracy in both one-page-out and half-page-out experiments. Surprisingly, for harmonic BLC, the performance is slightly superior in the half-pages-out than in the one-page-out experiments. This is due to the fact that for one page the performance is only 87.3%; the performance for all other pages is always above 97.2%, and is 100% for 23 pages in the dataset. The poor performance on one particular page drags down the average for one-page-out, compared to half-pages-out where better-performing pages ameliorate the average.

5 Conclusions and Future Work

The high accuracy achieved by both logistic regression and the harmonic BLC algorithm confirm our basic hypothesis: the set of users that interacts with news posts in social network sites can be used to predict whether posts are hoaxes.

We presented two techniques for exploiting this information: one based on logistic regression, the other on boolean label crowdsourcing (BLC). Both algorithms provide good performance, with the harmonic BLC algorithm providing accuracy above 99% even when trained over sets of posts consisting of 0.5% of the full dataset (or about 80 posts). This suggests

that the algorithms can scale up to the size of entire social networks, while requiring only a modest amount of manual classification.

We also analyzed the extent to which our performance depends on the community of users naturally aggregating around pages of similar content. We showed that the harmonic BLC algorithm can transfer information across pages: even when only half of the pages are represented in the training set, the performance is above 99%. Even on the "intersection dataset", consisting of only users who liked *both* hoax and non-hoax posts, our methods achieve performance of 90%, albeit requiring for this a training set consisting of 10% of the posts; this produces evidence that our approach might work even when applied to communities of users that are not strongly polarized towards scientific vs. conspiracy pages. We note that the intersection dataset is a borderline example that does not occur in the communities we studied. Together, these results seem to indicate that the techniques proposed may be sufficiently robust for an extensive application in a real-world scenario.

Future work involves the implementation of the presented method within a real-world Facebook online automated hoax detection system. To do this, two steps are foreseen: *i*) the extension to other community languages besides the Italian community considered as example application in this work, and *ii*) the classification of posts for the associated extension of the ground truth. For the first point, under the assumption that there is no substantial difference among countries and language communities, the method can be replicated by appropriately enlarging the ground truth to include posts (and therefore users) not related to the Italian Facebook community. For the second point, in this work we assumed that all posts published by conspiracy pages can be classified as hoaxes, and that all posts published by scientific pages can be classified as non-hoaxes. This merely practical simplification, based on the approach and findings in [3], avoided the need for a manual classification of the individual posts. However, in a real-world application, single post classification can of course be adopted. Additionally, we see the interest of evaluating the use of other machine learning methods besides logistic regression and harmonic crowdsourcing.

References

[1] P. Heymann, G. Koutrika, and H. Garcia-Molina. "Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges". In: *IEEE Internet Computing* 11.6 (Nov. 2007), pp. 36–45.

[2] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi. “Science vs Conspiracy: Collective Narratives in the Age of Misinformation”. In: *PLOS ONE* 10.2 (Feb. 2015), e0118093.

[3] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. “The Spreading of Misinformation Online”. en. In: *Proceedings of the National Academy of Sciences* 113.3 (Jan. 2016), pp. 554–559.

[4] J. C. Hernandez, C. J. Hernandez, J. M. Sierra, and A. Ribagorda. “A First Step towards Automatic Hoax Detection”. In: *Proceedings. 36th Annual 2002 International Carnahan Conference on Security Technology*. 2002, pp. 102–114.

[5] J. Mason. “Filtering spam with spamassassin”. In: *HEANet Annual Conference*. 2002, p. 103.

[6] T. Petković, Z. Kostanjčar, and P. Pale. *E-Mail System for Automatic Hoax Recognition*. 2005.

[7] A. Ishak, Y. Y. Chen, and S.-P. Yong. “Distance-Based Hoax Detection System”. In: *2012 International Conference on Computer Information Science (ICCIS)*. Vol. 1. June 2012, pp. 215–220.

[8] M. Vuković, K. Pripužić, and H. Belani. “An Intelligent Automatic Hoax Detection System”. en. In: *Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, Berlin, Heidelberg, Sept. 2009, pp. 318–325.

[9] I. Yevseyeva, V. Basto-Fernandes, D. Ruano-Ordás, and J. R. Méndez. “Optimising Anti-Spam Filters with Evolutionary Algorithms”. In: *Expert Systems with Applications* 40.10 (Aug. 2013), pp. 4010–4021.

[10] M. Sharifi, E. Fink, and J. G. Carbonell. “Detection of Internet Scam Using Logistic Regression”. In: *2011 IEEE International Conference on Systems, Man, and Cybernetics*. Oct. 2011, pp. 2168–2172.

[11] L. Mui, M. Mohtashemi, and A. Halberstadt. “A Computational Model of Trust and Reputation for E-Businesses”. In: *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 7 - Volume 7*. HICSS '02. Washington, DC, USA: IEEE Computer Society, 2002.

[12] C. Dellarocas. “The digitization of word of mouth: Promise and challenges of online feedback mechanisms”. In: *Management science* 49.10 (2003), pp. 1407–1424.

[13] J. Golbeck and J. Hendler. “Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks”. en. In: *Engineering Knowledge in the Age of the Semantic Web*. Springer, Berlin, Heidelberg, Oct. 2004, pp. 116–131.

[14] B. T. Adler and L. de Alfaro. “A Content-Driven Reputation System for the Wikipedia”. In: *Proceedings of the 16th International Conference on World Wide Web*. WWW '07. Banff, Alberta, Canada: ACM, 2007, pp. 261–270.

[15] M. Potthast, B. Stein, and R. Gerling. “Automatic vandalism detection in Wikipedia”. In: *European Conference on Information Retrieval*. Springer. 2008, pp. 663–668.

[16] B Adler, L. De Alfaro, S. Mola-Velasco, P. Rosso, and A. West. “Wikipedia vandalism detection: Combining natural language, metadata, and reputation features”. In: *Computational linguistics and intelligent text processing* (2011), pp. 277–288.

[17] S. Kumar, R. West, and J. Leskovec. “Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes”. In: *Proceedings of the 25th*

International Conference on World Wide Web. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 591–602.

[18] X. Chen, R. Chandramouli, and K. P. Subbalakshmi. "Scam Detection in Twitter". en. In: *Data Mining for Service*. Ed. by K. Yada. Studies in Big Data 3. Springer Berlin Heidelberg, 2014, pp. 133–150.

[19] J. Ito, J. Song, H. Toda, Y. Koike, and S. Oyama. "Assessment of Tweet Credibility with LDA Features". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. New York, NY, USA: ACM, 2015, pp. 953–958.

[20] D. R. Karger, S. Oh, and D. Shah. "Iterative Learning for Reliable Crowdsourcing Systems". In: *Advances in Neural Information Processing Systems*. 2011, pp. 1953–1961.

[21] Q. Liu, J. Peng, and A. T. Ihler. "Variational Inference for Crowdsourcing". In: *Advances in Neural Information Processing Systems*. 2012, pp. 692–700.

[22] L. de Alfaro, V. Polychronopoulos, and M. Shavlovsky. "Reliable Aggregation of Boolean Crowdsourced Tasks". In: *Third AAAI Conference on Human Computation and Crowdsourcing*. 2015.