A Gamified Approach to Naïve Bayes Classification: A Case Study for Newswires and Systematic Medical Reviews

Giorgio Maria Di Nunzio
Dept. of Information Engineering
University of Padua, Italy
dinunzio@dei.unipd.it

Maria Maistro
Dept. of Information Engineering
University of Padua, Italy
maistro@dei.unipd.it

Federica Vezzani
Dept. of Linguistic and Literary Studies
University of Padua, Italy
federica.vezzani@phd.unipd.it

ABSTRACT

Supervised machine learning algorithms require a set of labelled examples to be trained; however, the labelling process is a costly and time consuming task which is carried out by experts of the domain who label the dataset by means of an iterative process to filter out non-relevant objects of the dataset. In this paper, we describe a set of experiments that use gamification techniques to transform this labelling task into an interactive learning process where users can cooperate in order to achieve a common goal. To this end, first we use a geometrical interpretation of Naïve Bayes (NB) classifiers in order to create an intuitive visualization of the current state of the system and let the user change some of the parameters directly as part of a game. We apply this visualization technique to the classification of newswire and we report the results of the experiments conducted with different groups of people: PhD students, Master Degree students and general public. Then, we present a preliminary experiment of query rewriting for systematic reviews in a medical scenario, which makes use of gamification techniques to collect different formulation of the same query. Both the experiments show how the exploitation of gamification approaches help to engage the users in abstract tasks that might be hard to understand and/or boring to perform.

CCS CONCEPTS

• **Information systems** → *Multilingual and cross-lingual retrieval*; Crowdsourcing; • **Theory of computation** → *Active learning*; • **Applied computing** → *Language translation*; Health care information systems; • **Software and its engineering** → *Interactive games*;

KEYWORDS

Automatic Text Classification, Gamification, Crowdsourcing, eHealth

ACM Reference Format:
Giorgio Maria Di Nunzio, Maria Maistro, and Federica Vezzani. 2018. A Gamified Approach to Naïve Bayes Classification: A Case Study for Newswires and Systematic Medical Reviews. In *The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3191547

WWW'18 Compaion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04..
https://doi.org/10.1145/3184558.3191547

1 INTRODUCTION

The creation of a ground-truth, or golden standard, in machine learning is usually very expensive as it requires a manual labelling of the objects by experts in the field. In order to reduce the costs of this labelling phase, it is possible to use crowd-sourcing and interactive machine learning approaches [1] to annotate datasets at affordable costs [27]. One major challenge in motivating people to participate in these labelling tasks is to design a system that promotes and enables the formation of positive motivations towards work as well as fits the type of the activity. In this sense, there are two important concepts to take into account: interpretability and gamification.

1.1 Interpretability

Interpretability is a common desiderata in applications of machine learning pertaining to expert-driven fields (like the medical field, for example) where the users want to understand and validate the meaning of a model before even considering deploying it. One goal of interpretability is to demystify the machine learning 'black-box' for non-experts by creating algorithms that can inform, collaborate with, compete with, and understand users in real-world settings [20]. For example, a good predictor would certainly be useful for the case of a model returning critical decisions, like the effectiveness of a drug in its therapeutic use. Nevertheless, making a model that reveals the reasons why the drug would or would not work in specific cases would be much more meaningful and would enable the experts to design better therapeutic drugs in the future.[1]

Another issue is that the interpretability and the accuracy of a model are concurrent tasks. For instance, in order to interpret the prediction of a classifier, it is usually suggested to use a few number of variables or few examples, but if you need accuracy, you have to use a lot of variables and adequate training dataset. Féraud, one of the authors of a paper introducing an approach to explain neural network classification [19], proposes a solution with two models: one for prediction and another one for interpretation. Another way to tackle this problem is to look for near-optimal solutions after training the classifier only on a small subset of the available samples, like the work on binary classification problems by [6]. In April 2016, in a session with Ricardo Baeza Yates at Quora,[2] somebody asked the question about "How important is interpretability for a model in Machine Learning?". Professor Yoshua Bengio replied:

[1] https://goo.gl/jntQJU
[2] https://goo.gl/MWYJpJ

> Interpretability has been overblown. What we really need before using a model is some (statistical) reassurance about the general ability of the trained model (which is what test error and estimating uncertainty around it aims to do). That being said, I think we should do everything we can to figure out what is going on inside machine learning models, because it can help us debug them and figure out their limitations, thus build better models.

1.2 Gamification

Gamification is defined as “the use of game design elements in non-game contexts” [7]. For example, game elements, such as leaderboards or points, are used for purposes different from their normal expected employment and serve as a summary of users accomplishments [2]. Nowadays, gamification spreads through a wide range of disciplines and its applications are implemented in different areas. For instance, an increasingly common feature of online communities and social media sites is a mechanism for rewarding user achievements based on a system of badges and points. They have been employed in many domains, as for example, games for health [26], for education [23] and for enterprises [34].

The use of gamification in academic research areas has been introduced very recently and its potential is still to be explored and validated. Information Retrieval (IR) has lately dealt with gamification, as witnessed by the GamifIR in 2014, 2015 and 2016[3]. In [21], the authors describe the fundamental elements and mechanics of a game and provide an overview of possible applications of gamification to the IR process. In [32], approaches to properly gamify Web search are presented, i.e. making both the search of information and the scanning of results a more enjoyable activity.

1.3 Our proposal

In this paper, we present our current work on the geometrical interpretation of the Bayes' rule inspired from the idea of classification in Likelihood Spaces [33]. This visual approach has been recently proposed as an intuitive way to teach machine learning and optimize probabilistic classifiers [8, 11–13, 16]. We introduce a set of experiments where non-expert users have used this type of visualization to directly interact with a Naïve Bayes (NB) classifiers. Moreover, we present a preliminary experiment exploiting a gamification approach in order to collect different reformulations of the same query in a medical scenario. With these two experiments we aim at showing how gamification may be helpful to collect human annotated data in different settings and with different approaches.

In Section 2, we introduce some basic concept of Interactive Machine Learning that are used to present the two-dimensional visual interpretation of a NB classifier. Successively, Section 3 describes the experiments on the use of gamification approaches for the problem of classification of news and medical documents. Finally, in Section 4, we discuss some open questions and give our final remarks.

2 INTERACTIVE MACHINE LEARNING

In Interactive Machine Learning (IML) the interaction with users allows models to be updated fast and very accurately; in addition, even non-expert users can solve machine learning problems with minimum effort by means of intuitive visualization tools [1].

In this context, Becker proposed a list of desired requirements for the visualization of the structure of classifiers [4]:

- to quickly grasp the primary factors influencing the classification with very little knowledge of statistics;
- to see the whole model and understand how it applies to records, rather than the visualization being specific to every record;
- to compare the relative evidence contributed by every value of every attribute;
- to see a characterization of a given class, that is a list of attributes that differentiate that class from others;
- to infer record counts and confidence in the shown probabilities so that the reliability of the classifier's prediction for specific values can be assessed quickly from the graphics;
- to interact with the visualization to perform classification;
- to have a system that should handle many attributes without creating an incomprehensible visualization or a scene that is impractical to manipulate.

Inspired by these requirements, which are still very relevant today, we focus on the problem of exploration and classification of large datasets by lay people. By providing adequate data and knowledge visualizations, users have a deeper understanding of the resulting classifier, and the pattern recognition capabilities of the user can be used to increase the effectiveness of the classifier construction [3, 10].

In the following section, we start by providing the principal concepts and notions that lay the basis for a NB classifier and then we describe the proposed interactive visualization of the NB classifier.

2.1 Two dimensional NB Classification

Different Bayesian approaches has been proposed to produce predictive models, which are not only accurate, but also interpretable by human experts. One example is the Bayesian Rule List, a model consisting of a series of if-then-statements that discretize a high-dimensional multivariate feature space, into a series of simple, readily interpretable, decision statements [25]. Another example, namely the Bayesian Case Model (BCM), is a general framework for Bayesian case-based reasoning and prototype classification and clustering. Experiments with users showed statistically significant improvements to participants' understanding when using explanations produced by BCM, compared to those given by prior state-of-the-art approaches [24].

In this paper we use an extension of Likelihood Spaces [33]. In particular, we consider the problem of binary classification defined as follows: suppose to work with a set of n classes $C = \{c_1, ..., c_i, ..., c_n\}$, and that an object o can be assigned to one (or more) than one class. Instead of building one single multi-class classifier, we split this multi-class categorization into n binary problems [31]. Usually, the two classes of a binary problem are: c_i, the ‘positive' class, and $\overline{c}_i = C \setminus c_i$ the ‘negative' class (we will drop the

[3] http://gamifir.com

index i when there is no risk of misinterpreting the formula). The simplest approach of a Bayesian classifier is to assign the object o to the positive category when

$$P(c|o) > P(\bar{c}|o) \tag{1}$$

that is, if the probability of the class c is greater than the probability of its complement $\bar{c}$ given the object o. Bayes' rule tells how we can reverse this problem using the prior probability $P(c)$ and the likelihood of the object $P(o|c)$:

$$\frac{P(o|c)P(c)}{P(o)} > \frac{P(o|\bar{c})P(\bar{c})}{P(o)} \tag{2}$$

The following step in the Likelihood Spaces projection is simply computing the log-likelihood of Equation (2) and consider $\log(P(o|c))$ and $\log(P(o|\bar{c}))$ as coordinates of a two-dimensional space where o is classified under category c when

$$\log(P(o|c)) - \log(P(o|\bar{c})) > \log(P(\bar{c})) - \log(P(c)) \tag{3}$$

In real case scenarios, we estimate the likelihood function by means of the class conditional probability of the features of the object o. For example, let us assume that the objects we want to study are characterized by a set features $\mathcal{F} = \{f_1, \ldots, f_m\}$. An object o is, therefore, a particular realization of these features, and its likelihood for category c is:

$$P(o|c) = P(\{f_1, \ldots, f_m\}|c) \tag{4}$$

An issue related to the estimation of this probability is that the amount of data needed grows exponentially with the number of features (e.g. if the variables in $\mathcal{F}$ are binary, the probability table has $2^{|\mathcal{F}|}$ entries [22]). For this reason, it is very common to simplify the problem by means of a strong assumption named *Naïve Bayes* assumption, i.e. all the features are conditionally independent given the class. In mathematical terms:

$$P(\{f_1, \ldots, f_m\}|c) = \prod_{j=1}^{m} P(f_j|c) \tag{5}$$

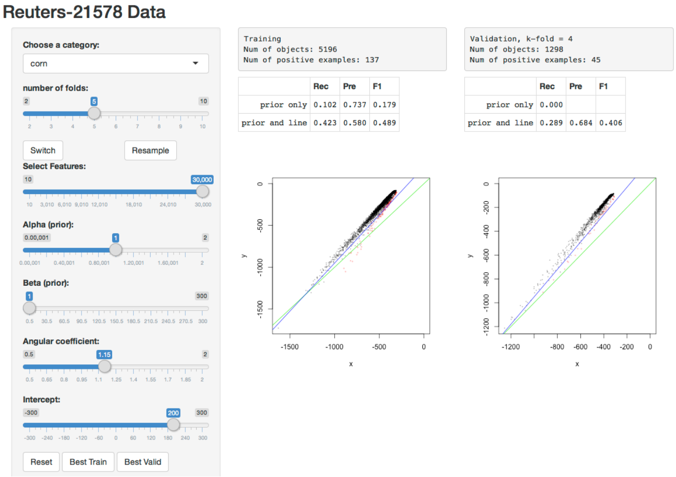
Then, the decision in the Likelihood Spaces becomes:

$$\sum_j \log(P(f_j|c)) - \sum_j \log(P(f_j|\bar{c})) > \log(P(\bar{c})) - \log(P(c)) \tag{6}$$

The Likelihood Spaces approach has been developed under the assumption of a zero-one loss function (equal unitary cost for both a false positive and a false negative). In [9, 17], we extended Equation (1) to a more general case that takes into account two more parameters:

$$\underbrace{P(o|\bar{c})}_{y} < m_L \underbrace{P(o|c)}_{x} + q_L \tag{7}$$

where m_L and q_L can be either set automatically, for example by optimizing a measure of classification accuracy, or semi-automatically by asking to a user to suggest the initial conditions based on a visual inspection of the problem. The interactive visual result of this extension is shown in Figure 1.[4] This Web application was implemented with the Shiny package in R [5]. and allows users to train a NB text classifier on a standard text collection, the Reuters-21578. [5] The two plots shown in Figure 1 illustrate the visualization of the likelihood space, with the x axis representing $P(o|c)$ and the y axis representing $P(o|\bar{c})$. Each point on the graph corresponds to an object (i.e. a document of the Reuters collection) and its color denotes the class to which the point belongs. The goal is very simple: finding the line that separates the two sets of points (the positive and the negative category) in the best possible way. According to the theory, we can improve the separation of the points in two ways: we can change the estimates of the probability of the features by modifying the values α and β of the prior beta function; i.e. we can adjust the classification line by changing the intercept q_L and the angular coefficient m_L in the Likelihood Spaces (see [9] for more details about this approach).



Reuters-21578 Data

Choose a category: corn

number of folds: 5

Select Features: 10 – 30,000

Alpha (prior): 1

Beta (prior): 1

Angular coefficient: 1.15

Intercept: 200

```
Training
Num of objects: 5196
Num of positive examples: 137
```

	Rec	Pre	F1
prior only	0.102	0.737	0.179
prior and line	0.423	0.580	0.489

```
Validation, k-fold = 4
Num of objects: 1298
Num of positive examples: 45
```

	Rec	Pre	F1
prior only	0.000		
prior and line	0.289	0.684	0.406

Figure 1: Two-dimensional view of a NB classifier.

3 GAMIFICATION FOR NEWSWIRES AND MEDICAL SYSTEMATIC REVIEWS

In this section, we present a set of experiments describing the refinements of the probabilistic text classifier visualization approach which was transformed into a game. The game is based on the two-dimensional representation of probabilities presented in Equation (7). Moreover, we present a preliminary experiment of query rewriting in the context of medical systematic reviews.

3.1 Gamified Classification of Newswires

The problem of classification of newswires with a gamified approach has been tested with different users and different game interfaces, that were more and more simplified compared to the original one.

The initial version of the interface[6], shown in Figure 1, was designed to be used by experts to understand how to optimize the search of the optimal parameters. In the "gamified" version of this problem, players have to find the best combination of m_L and q_L having a fixed amount of resources available to train and validate the algorithm. The game is organized in N levels (corresponding to the binary classification problems), which are presented from the easiest to the most difficult and correspond to the different

[4] https://gmdn.shinyapps.io/shinyK/

[5] http://www.daviddlewis.com/resources/testcollections/

[6] Available at https://gmdn.shinyapps.io/shinyK/

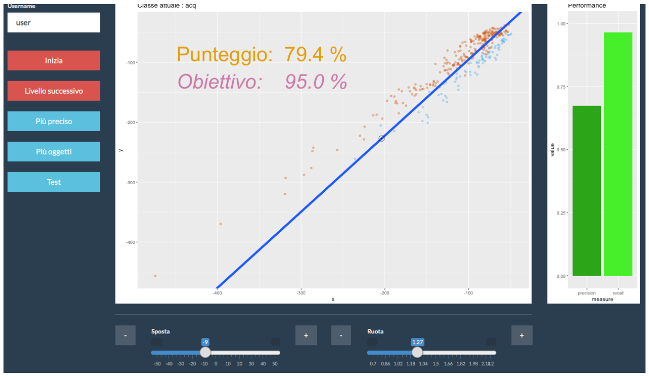

Punteggio: 79.4 %
Obiettivo: 95.0 %

Figure 2: Layout for students

classification tasks of the top N classes of the Reuters 21578 dataset. A level is difficult when it is hard to linearly separate the positive class c and the negative class $\bar{c}$. An object can be used during the game either as a training example or as validation sample, but not both. The goal of each level (and in general of the game) is to find the best classifier, i.e. the one which maximizes the F1 score, with the least amount of resources. Resources can be used to increase the number of objects of the training and/or the validation set. At any point in the game, the player can use some resources to buy additional training or validation objects. By doing so, an additional 5% of the collection is added to either the training set (more precise) or validation set (more objects on the screen). Once the player has found what he/she considers the best classifier, he/she can proceed with the test, thus the classifier is tested on the test set and the F1 score is computed. At this point, the level is completed and the player is forced to go to the next level or conclude the game.

A second version of the interface was designed for PhD and post-doc students[7] and a pilot study was carried out to test this preliminary version of the game and to collect opinions and suggestions regarding possible improvements of the game [15, 28].

During the European Researcher's Night at the University of Padua in September 2016, we designed a third version of the interface to make the game easier for kids of primary and secondary schools, who played with the application [29]. The interface, shown in Figure 2, lets users play only three "levels" (the levels are the categories) and gives feedback about the current performance whenever the line is adjusted. In this experiment, we also added some incentives like a public leaderboard, that was displayed and regularly updated, and chocolate candies for the top scorer.

The first week of April 2017, during an event for the brand new 50 euro note, located at one of the branches of Banca d'Italia in Padua, we presented a fourth version of the game that was available for the public for a whole week. For this study [29], we decided to make the layout cleaner, see Figure 3, and add keyboard controls to change the decision line instead of using sliders. We kept the same game incentives, chocolate candies and leaderboard, and we added an instructional presentation of the problem to help the player to understand what 'machine learning' and 'training set' are.

[7] Available at https://gmdn.shinyapps.io/Classification/

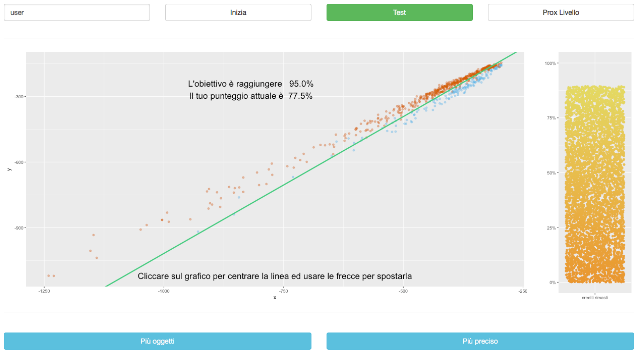

L'obiettivo è raggiungere 95.0%
Il tuo punteggio attuale è 77.5%
Cliccare sul grafico per centrare la linea ed usare le frecce per spostarla

Figure 3: Layout for general public

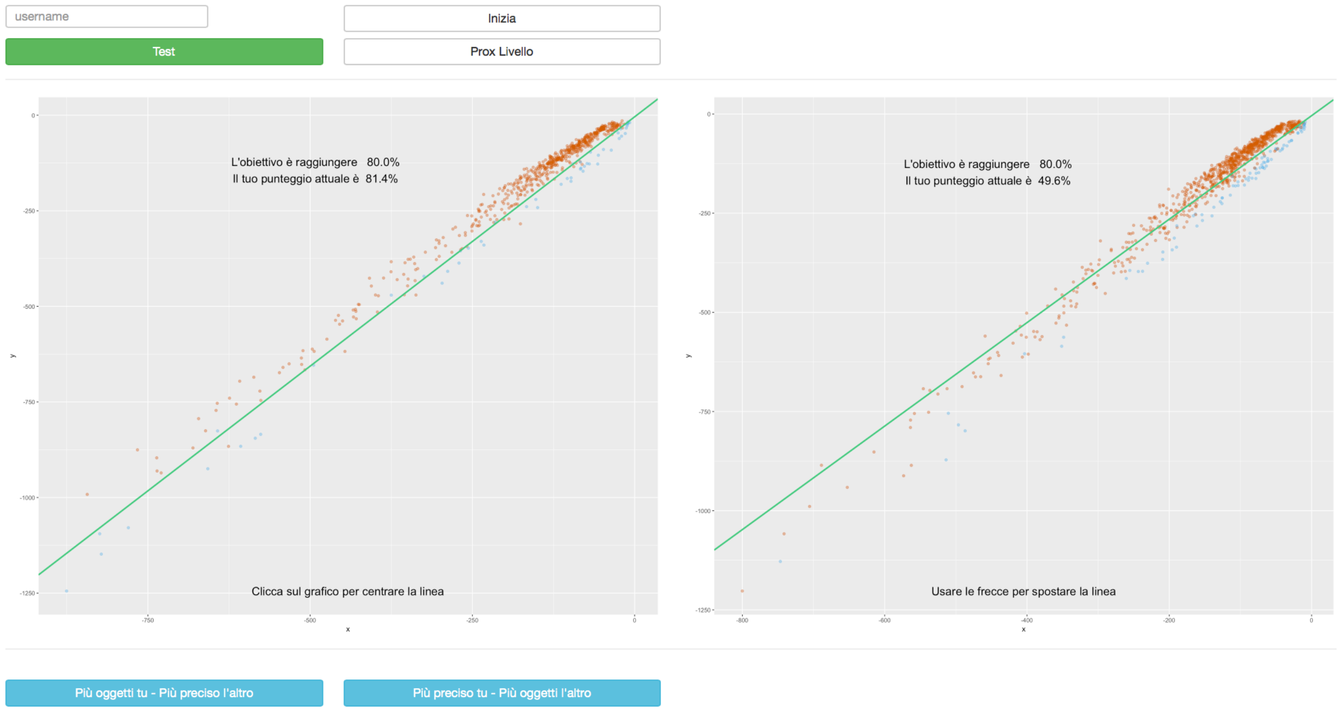
Table 1: Manual vs NB and SVM classifiers. Classification performance during the European Researcher's Night. The averaged F1 measure of 28 participants is reported for each class.

Level	Goal	Manual	NB	SVM
1	0.950	0.931	0.943	0.940
2	0.850	0.784	0.768	0.840
3	0.750	0.715	0.715	0.730
average	0.850	0.810	0.809	0.837

Finally, a fourth and last version of the interface was designed for the European Researcher's night, organized by the University of Padua in September 2017. Analogously to the previous experiments, the game evolved through the same three levels: the users could see the current score and goal when the line was adjusted, and they could decide to spend some resources to increase the number of training and validation objects. The main difference between the previous versions of the game was the introduction of a collaborative component into the game dynamic. As shown in Figure 4, the screen was divided in two separate canvases, each one representing the visualization of the objects and the classifier. The player on the left could control the slope of the line, i.e. the m_L parameter, while the player on the right could set the intercept height, i.e. q_L. The players needed to collaborate in order to decide the best location of the line and how much resources to spend. Moreover, the points displayed on the left plot were different from those displayed on the right plot. Therefore players needed to find a compromise to place the classification line in order to maximize both the F1 scores. Even for this experiment, we added chocolate candies as an incentive to achieve the best performance using, at the same time, as less resources as possible.

3.2 Evaluation of the Classification Game

In this section we present some experimental results corresponding to the data collected during the European Researcher's Night in 2016 shown in Table 1, the event for the new 50 euro note shown in Table 2, and the European Researcher's Night in 2017 shown in Table 3. All the experiments are conducted with the Reuters-21578 text collection from which we selected the following three



L'obiettivo è raggiungere 80.0%
Il tuo punteggio attuale è 81.4%

Clicca sul grafico per centrare la linea

L'obiettivo è raggiungere 80.0%
Il tuo punteggio attuale è 49.6%

Usare le frecce per spostare la linea

Figure 4: Layout for a collaborative two-player game. The player on the left can buy resources (training and validation objects) while the player on the right can shift and rotate both lines. The points that a player uses as validation points (those points that are actually on the plot) are used as training points for the other player and viceversa, like a sort of 2-fold cross validation approach.

Table 2: Manual vs NB and SVM classifiers. Classification performance during the week at the Banca d'Italia. The averaged F1 measure of 27 participants is reported for each class.

Level	Goal	Manual	NB	SVM
1	0.950	0.940	0.942	0.939
2	0.850	0.807	0.786	0.841
3	0.750	0.714	0.710	0.723
average	0.850	0.830	0.813	0.834

Table 3: Manual vs NB and SVM classifiers. Classification performance during the European Researchers night in September 2017. The averaged F1 measure of 14 **round is reported for each level.**

Level	Goal	Manual	NB	SVM
1	0.950	0.831	0.939	0.942
2	0.850	0.805	0.787	0.845
3	0.750	0.731	0.713	0.716
average	0.850	0.789	0.813	0.834

categories: *acq* (level 1), *crude* (level 2), *money-fx* (level 3). Each table report the F1 scores of the gold system, which is the score obtained with a NB classifier trained with the whole validation and training set and with the optimal parameters m_L and q_L found by an automatic approach [9], as well as the performances of the classifier with the human interaction, referred as manual, and a NB and a Support Vector Machine (SVM) classifiers, trained on the same data points used by players. For each level the results are averaged on the number of participants per level. Notice that, since each participant was using a random subset of the text collection, the algorithms were trained on a different amount of data during the game, therefore the scores in Table 1, Table 2, and Table 3 are not directly comparable.

A total of 28 players used the interface during the European Researcher's Night in 2016. Table 1 presents their results showing that the manual approach reaches performances very close the NB classifier and sometimes it performs even better. The results in terms of classification performance are quite surprising if you consider that these users were mainly children who did not know anything about machine learning or text classification.

During the presentation of the brand new 50 euro note a total of 27 participants played with the game. Their results, reported in Table 2, are similar to those obtained from the previous experiment. The interaction of the users with the algorithm through the gamified approach reached performances close to SVM and sometimes better than NB. The amount of resources used was comparable to the experiment European Researcher's Night: players tend to consider the performance of the classifier satisfactory when 30% of the resources are used.

Finally, Table 3 shows the results obtained during the European Researcher's Night in 2017. We collected the data from 14 rounds, which corresponds to 28 players. The average performance of the manual classifier is worse than the performances of NB and SVM. However, considering the partial scores segmented by levels, you can notice that the manual classifier performs better than NB on level 2 and it performs better than both NB and SVM on level 3. This is consistent with the results of Table 1 and Table 2, where the manual classifier perform better than NB on level 2 and 3. Remind that level 3 is the hardest, i.e. it represents one of the class whose points are most difficult to separate through a straight line. Even if these results are preliminary and further work is needed to consolidate them, they suggest that involving the human in the classification process might be more helpful when the problem is more difficult, i.e. when the positive and negative classes are partially overlapped.

To conclude, we want to remark that we are not claiming that a human by simply tweaking the classifier parameters can perform better than a NB classifier trained on the whole training and validation sets with an optimal strategy to tune the parameters. Indeed, the experimental results show that this is not true, and the manual approach always gains lower F1 scores than the gold. However, we want to highlight that the classifier tuned by humans performs similarly and sometimes even better than NB and SVM while only using around the 30% of the resources.

3.3 Interactive Systematic Medical Reviews

In this section, we illustrate a gamified experiment of query rewriting in the context of medical eHealth, which was performed with the students of the Master's Degree course in Modern Languages for International Communication and Cooperation of the University of Padua. In particular, we re-proposed a task previously performed with our participation in the Cross-Language Evaluation Forum (CLEF) eHealth Task 2: "Technologically Assisted Reviews in Empirical Medicine" [18]. The task consisted in retrieving all the relevant documents for medical specific domains as early as possible and with the least effort.

The query rewriting experiment was presented in a semi-interactive gamified setting, which was appreciated and positively evaluated by those students who completed the task. The initial idea behind this experiment was to make available to non-expert users in the medical field, i.e. the students of the master's degree course, an interactive system allowing to enter the query reformulations and which gives automatically a feedback of relevance on the documents appearing from their research. Users can reformulate iteratively their queries in order to obtain more relevant documents for a specific topic through aspects of gamification (such as score, ranking etc) [35]. Due to time and course organization issues, the system was partially implemented at the time of the experiment; we therefore proceeded by presenting this experiment as a "role-playing game" with three main characters: the physician, represented by the professor, the project manager of a translation agency, who was a PhD student, and the translators, interpreted by the students of the Translation technologies' class. The physician asked to the project manager the translations of the abstracts of the most relevant documents for a specific topic. The project manager entrusted all the in-house translators with the request in order to satisfy the needs of the commissioner and provided the participants with all the information and, in particular, the methodology that underlies the experiment.

In order to retrieve all the relevant documents for the specific medical domains, participants were instructed to proceed with the following strategy: reformulate an initial query given by the domain expert (i.e. the physician), with different levels of specificity by performing the analysis of some linguistic and terminological aspects. In addition, the students could exploit the information given by the provided documents, that could be relevant or not according to the initial query, the list of term frequencies, documents frequencies, and the boolean query generated by PubMed.[8]

The first variant of the query was a list of keywords that the participants obtained from the semic analysis [30] of the technical terms contained in the initial query in order to cover as much as possible the semantic sphere affected by the term analyzed. The second variant was instead a human readable reformulation, therefore grammatically correct, and containing the fewest possible number of terms equal to the starting query. This reformulation is therefore made up of synonymic variants, acronyms, abbreviations or periphrases. The third variant was different from the previously proposed and it does not follow any precise approach other than that of human interpretation resulting from the approximate study of the subject contained in the query.

With this kind of analysis participants were able to

(1) create the basis of knowledge for the domain and the context of study;
(2) propose the query variant through three different approaches.

The experiment involved 90 students, all of them with different backgrounds, who were divided into 30 groups of 3 people each. The physician had 30 "information needs" to satisfy, thus each group was entrusted with one specific information need from the medical field. 28 groups completed the task and we received a total of 28 list of keywords (first version), 28 human-readable reformulation (second version) and 66 individual reformulations (third version). Hereinafter an example of the three variants proposed by a group of students for a specific information need is given:

- Initial query: Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain;
- First variant: Sensitivity, specificity, test, tests, diagnosis, examination, physical, straight leg raising, slump, radicular, radiculopathy, pain, inflammation, compression, compress, spinal nerve, spine, cervical, root, roots, sciatica, vertebrae, lumbago, LBP, lumbar, low, back, sacral, disc, discs, disk, disks, herniation, hernia, herniated, intervertebral;
- Second variant: Sensitivity and specificity of physical tests for the diagnosis of nerve irritation caused by damage to the discs between the vertebrae in patients presenting LBP (lumbago)
- Individual reformulation: Patients with pain in the lower back need a check-up for the compression or inflammation

[8] https://www.ncbi.nlm.nih.gov/pubmed/

of a spinal nerve caused by rupture of fibrocartilagenous material that surrounds the intervertebral disk.

This approach has contributed to an effective and efficient reformulation for the retrieval of the most relevant documents for the creation of systematic reviews [14]; it also has produced a set of terminological records following the model implemented in an eHealth linguistic resource: TriMED [36]. A preliminary analysis of the results has also shown that the the stimulus provided by the fact that this experiment was part of a study aiming to compare linguistic experts with students was sufficient to motivate students to work and have fun at the same time. We believe that a fully gamified experience will motivate users even more allowing to collect more higher quality data.

4 DISCUSSION AND FINAL REMARKS

In this paper, we presented two different gamification approaches, a first approach applied to a NB classifier with newswire, and a second approach applied to the medical domain. To visualize and interact with the NB classifier we exploited a geometrical interpretation of the Likelihood Spaces on a two dimensional spaces, where the objects are represented by points on the space and the classifier is a straight line separating them. We successively refined this visualization to develop more and more clean and easy versions, suitable to be used with kids and the general public, who do not have any understanding of machine leaning and NB classifier. Preliminary experimental results are promising and shows that the integration of the human in the classification process, particularly in the placement of the decision line, can be beneficial for classification performances. Moreover, we proposed a second gamification approach applied to the medical domain. These new set of experiments have been used to study an early stopping strategy for systematic medical reviews and to create a set of high quality multilingual terminological records.

Future directions will extend the proposed interface in order to account for the labelling process. The user will be able to label some extra documents in order to increase the performance of the classifier. Moreover, these documents will not be randomly chosen, but they will be selected with active learning strategies thus providing valuable additional information for the classification problem. Finally, the proposed interface might be adapted for an interactive query rewriting approach which can be used by both experts (physicians) and non-experts (patients) to find multilingual medical information. As previously mentioned, the idea is to design an interface which suggests the amount of relevance information that is still missing allowing users to improve their own reformulations. In addition, the interface will also provide alternative terms for the query in order to interactively suggest a proper terminology for the reformulation of the information the user is looking for.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (2014), 105–120. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2513

[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering User Behavior with Badges. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 95–106. https://doi.org/10.1145/2488388.2488398

[3] Mihael Ankerst, Martin Ester, and Hans-Peter Kriegel. 2000. Towards an Effective Cooperation of the User and the Computer for Classification. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00)*. ACM, New York, NY, USA, 179–188. https://doi.org/10.1145/347090.347124

[4] Barry G. Becker. 1997. Using MineSet for Knowledge Discovery. *IEEE Computer Graphics and Applications* 17, 4 (1997), 75–78.

[5] Winston Chang. 2015. *Shiny: Web Application Framework for R.* http://CRAN.R-project.org/package=shiny R package version 0.11.

[6] Sanjeeb Dash, Dmitry M. Malioutov, and Kush R. Varshney. 2015. Learning interpretable classification rules using sequential rowsampling. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015.* 3337–3341. https://doi.org/10.1109/ICASSP.2015.7178589

[7] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proc. of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek '11)*. ACM, New York, NY, USA, 9–15. https://doi.org/10.1145/2181037.2181040

[8] Emanuele Di Buccio and Giorgio Maria Di Nunzio. 2013. A Visual Analysis of the Effects of Assumptions of Classical Probabilistic Models. In *International Conference on the Theory of Information Retrieval, ICTIR '13, Copenhagen, Denmark, September 29 - October 02, 2013.* 28. https://doi.org/10.1145/2499178.2499200

[9] Giorgio Maria Di Nunzio. 2014. A new decision to take for cost-sensitive Naïve Bayes classifiers. *Inf. Process. Manage.* 50, 5 (2014), 653–674. https://doi.org/10.1016/j.ipm.2014.04.008

[10] Giorgio Maria Di Nunzio. 2014. Visual Classification. In *Data Classification: Algorithms and Applications*, Charu C. Aggarwal (Ed.). CRC Press, Chapter 23, 607–632. http://www.crcnetbase.com/doi/abs/10.1201/b17320-24

[11] Giorgio Maria Di Nunzio. 2015. Shiny on Your Crazy Diagonal. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015.* 1031–1032. https://doi.org/10.1145/2766462.2767867

[12] Giorgio Maria Di Nunzio. 2015. Teaching Machine Learning: A Geometric View of Naïve Bayes. In *Research and Advanced Technology for Digital Libraries - 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015. Proceedings.* 343–346. https://doi.org/10.1007/978-3-319-24592-8_31

[13] Giorgio Maria Di Nunzio. 2016. Can you learn it? Probably! Developing Learning Analytics Tools in R. In *Joint Conference on Digital Libraries (JCDL 2016), Newark, New Jersey, USA.* ACM, In press.

[14] Giorgio Maria Di Nunzio. 2018 (in press). A Study of an Automatic Stopping Strategy for Technologically Assisted Medical Reviews. In *Advances in Information Retrieval. Proc. 40th European Conference on IR Research (ECIR 2018)*, Springer (Ed.).

[15] Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. 2016. Gamification for Machine Learning: The Classification Game. In *Proceedings of the Third International Workshop on Gamification for Information Retrieval co-located with 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016), Pisa, Italy, July 21, 2016.* 45–52. http://ceur-ws.org/Vol-1642/paper7.pdf

[16] Giorgio Maria Di Nunzio and Alessandro Sordoni. 2012. A visual tool for bayesian data analysis: the impact of smoothing on naive bayes text classifiers. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012.* 1002. https://doi.org/10.1145/2348283.2348427

[17] Giorgio Maria Di Nunzio. 2017. *Interactive Text Categorisation: The Geometry of Likelihood Spaces.* Springer International Publishing, Cham, 13–34. https://doi.org/10.1007/978-3-319-46135-9_2

[18] Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève Henrot. 2017. An Interactive Two-Dimensional Approach to Query Aspects Rewriting in Systematic Reviews. IMS Unipd At CLEF eHealth Task 2. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.* http://ceur-ws.org/Vol-1866/paper_119.pdf

[19] Raphael Féraud and Fabrice Clérot. 2002. A Methodology to Explain Neural Network Classification. *Neural Netw.* 15, 2 (March 2002), 237–246. https://doi.org/10.1016/S0893-6080(01)00127-7

[20] Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello (Eds.). 2016. *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings.* Lecture Notes in Computer Science, Vol. 9626. Springer. https://doi.org/10.1007/978-3-319-30671-1

[21] Luca Galli, Piero Fraternali, and Alessandro Bozzon. 2014. On the Application of Game Mechanics in Information Retrieval. In *Proc. of the 1st Int. Workshop on Gamification for Information Retrieval (GamifIR'14)*. ACM, New York, NY, USA, 7–11. https://doi.org/10.1145/2594776.2594778

[22] T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer. http://books.google.it/books?id=tVIjmNS3Ob8C

[23] Karl M Kapp. 2012. *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education.* John Wiley & Sons.

[24] Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1952–1960.

[25] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.* 9, 3 (09 2015), 1350–1371. https://doi.org/10.1214/15-AOAS848

[26] Simon McCallum. 2012. Gamification and Serious Games for Personalized Health. *Studies in Health Technology and Informatics* 177, 2012 (2012), 85–96.

[27] B. Morschheuser, J. Hamari, and J. Koivisto. 2016. Gamification in Crowdsourcing: A Review. In *2016 49th Hawaii International Conference on System Sciences (HICSS).* 4375–4384. https://doi.org/10.1109/HICSS.2016.543

[28] Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. 2016. Gamification for IR: The Query Aspects Game. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA2016), Napoli, Italy, December 5-7, 2016. (CEUR Workshop Proceedings)*, Pierpaolo Basile, Anna Corazza, Francesco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli (Eds.), Vol. 1749. CEUR-WS.org. http://ceur-ws.org/Vol-1749/paper21.pdf

[29] Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. 2017. A Game of Lines: Developing Game Mechanics for Text Classification. In *Proceedings of the 8th Italian Information Retrieval Workshop, Lugano, Switzerland, June 05-07, 2017. (CEUR Workshop Proceedings)*, Fabio Crestani, Tommaso Di Noia, and Raffaele Perego (Eds.), Vol. 1911. CEUR-WS.org, 40–47. http://ceur-ws.org/Vol-1911/7.pdf

[30] F. Rastier. 1987. *Sémantique interprétative.* Presses universitaires de France. https://books.google.it/books?id=BnuyAAAAIAAJ

[31] Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* 34, 1 (March 2002), 1–47. https://doi.org/10.1145/505282.505283

[32] Mark Shovman. 2014. The Game of Search: What is the Fun in That?. In *Proc. of the 1st Int. Workshop on Gamification for Information Retrieval (GamifIR'14).* ACM, New York, NY, USA, 46–48. https://doi.org/10.1145/2594776.2594786

[33] Rita Singh and Bhiksha Raj. 2004. Classification in Likelihood Spaces. *Technometrics* 46, 3 (2004), 318–329. https://doi.org/10.1198/004017004000000347 arXiv:http://www.tandfonline.com/doi/pdf/10.1198/004017004000000347

[34] Jennifer Thom, David Millen, and Joan DiMicco. [n. d.]. Removing Gamification from an Enterprise SNS.

[35] Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. 2016. ScentBar: A Query Suggestion Interface Visualizing the Amount of Missed Relevant Information for Intrinsically Diverse Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16).* ACM, New York, NY, USA, 405–414.

[36] Federica Vezzani, Giorgio Maria Di Nunzio, and Geneviève Henrot. 2018. In press.. TriMED: A Multilingual Terminological Database. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation LREC 2018, Miyazaky, Japan, May 7-12, 2018.*