Using Fuzzy Fingerprints for Cyberbullying Detection in Social Networks

Hugo Rosa
INESC-ID
Instituto Superior Técnico, Universidade de Lisboa
Faculdade de Psicologia, Universidade de Lisboa
Lisboa, Portugal
hugo.rosa@l2f.inesc-id.pt

Pável Calado
INESC-ID
Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal
pavel.calado@tecnico.ulisboa.pt

Ricardo Ribeiro
INESC-ID
ISCTE-IUL, Instituto Universitário de Lisboa
Lisboa, Portugal
ricardo.ribeiro@inesc-id.pt

Joao P. Carvalho
INESC-ID
Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal
joao.carvalho@inesc-id.pt

Bruno Martins
INESC-ID
Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal
bruno.g.martins@tecnico.ulisboa.pt

Luisa Coheur
INESC-ID
Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal
luisa.coheur@inesc-id.pt

***Abstract*—As cyberbullying becomes more and more frequent in social networks, automatically detecting it and pro-actively acting upon it becomes of the utmost importance. In this work, we study how a recent technique with proven success in similar tasks, Fuzzy Fingerprints, performs when detecting textual cyberbullying in social networks. Despite being commonly treated as binary classification task, we argue that this is in fact a retrieval problem where the only relevant performance is that of retrieving cyberbullying interactions. Experiments show that the Fuzzy Fingerprints slightly outperforms baseline classifiers when tested in a close to real life scenario, where cyberbullying instances are rarer than those without cyberbullying.**

***Index Terms*—cyberbullying, fuzzy fingerprints, machine learning, abusive language**

I. Introduction

Cyberbullying, a form of bullying or harassment using electronic means, has become a worldwide concern due to the increasingly widespread means to access social networks by children and teenagers. A 2016 UNESCO report [1] highlighted this issue by presenting case studies and policies from several distant countries, both geographically and culturally. From Southern Africa to East Asia, the Arab Region, Lithuania, Finland and the United States, the report shows that cyberbullying is not only a matter of public health, but also human rights. As a phenomenon, cyberbullying is defined as an individual's intentional, deliberate and repeated acts of cruelty to others through harmful posts, messages, or other forms of social aggression through various digital technologies [2]. Although it is a variation of real life bullying, the fact that it happens online means it has its own specificities. For instance, once a cyberbullying act is committed, it may remain publicly online forever, not only breaking any physical barriers required for bullying [3], but also creating the possibility for the victim to experience the same act repeatedly and indefinitely [4]. Additionally, an aggressor may retain its anonymity [5] and does not need to be physically stronger than the victim in order to torment him/her [6]. However, the fact that the aggression occurs online also provides victims with defense mechanisms, such as the ability of being offline or, depending on the social network, report improper behavior. Jang et al. [7] has even documented cases of bullying victims that become the aggressor in cyberbullying because it can be a form of retaliation towards their own aggressor. The consequences of cyberbullying are also well documented. In teenage victims, it can range from social exclusion to depression, poor academic results and, in extreme situations, suicide [8]. Despite the overwhelming acknowledgment of cyberbullying as an issue, it remains an open unsolved problem, since from a Machine Learning (ML) viewpoint it is usually addressed as a binary classification problem where the heavy unbalancing of the phenomenon is largely ignored. Yet, it is undeniable that one class is thematically broad and numerically large (the no-cyberbullying class - NCB), and the other, the cyberbullying class - CB, is very specific, has a much smaller number of instances, and hence, should be the only relevant one in what concerns the performance of any "Cyberbullying detection system". In this work, we present and analyze results from

978-1-5090-6020-7/18/$31.00 ©2018 IEEE

a set of experiments that part with the current trend in textual cyberbullying detection within the supervised machine learning community. We focus on the capability to detect/retrieve cyberbullying instances in unbalanced datasets and compare how Fuzzy Fingerprints (FFP) perform in this task against common ML approaches. This paper is organized as follows: In Section 2, we introduce some related work; In Section 3, we present the Fuzzy Fingerprints algorithm and how it has been successfully used in other text-based classification problems; In Section 4 the experimental setup is detailed and results are analyzed in Section 5; Concluding remarks are provided in Section 6.

II. Related Work

Although Information Retrieval (IR) and Natural Language Processing (NLP) techniques have been successfully applied to several text-based classification problems (sentiment analysis, topic detection, machine translation, text summarization, etc.), their application to cyberbullying detection is rather recent. In 2011, Bayzick et al. [9] built a very basic program based on a dictionary of keywords to detect cyberbullying in online conversations, with 58.6% accuracy. In the same year, Reynolds et al. [10] collected and labeled data from Formspring.me and were able to correctly identify 78.5% of the posts containing cyberbullying (recall), while Dinakar et al. [11] focused on detecting sensitive topics within cyberbullying annotated YouTube comments and found that they are mostly racially motivated, sexual harassment, or attacks to one's intelligence.

The most common approach to cyberbullying detection is through feature engineering, which expands the usual bag-of-words representation of text, by creating additional features that use domain knowledge of the data and attempt to improve the classifiers performance. Davdar et al. [12] improved an SVM's cyberbullying detection performance on YouTube comments by adding user features such as the frequency of profanity, number of pronouns, length of comments, usage of capital letters and use of emoticons. Pronoun counts are also used as features by Chavan et al. [13], which, alongside n-grams and skip-grams representations, improve performance and achieved a precision of 0.77 using a SVM classifier. Huang et al. [14] combined textual features (density of bad words, Part-Of-Speech tagging) and social networks features (by representing messages in a graph) to classify cyberbullying in a Twitter dataset. Similarly, Nahar et al. [15] propose the creation of a cyberbullying weighted directed graph model that can be used to compute predator and victim scores of each user, while using a weighted tf-idf scheme with textual features (second person pronouns and foul words) to enhance cyberbullying detection. Dinakar et al. [16] expanded on their previous work, creating label specific features (racial, sexual, intelligence) and also attempting to classify more indirect and implicit stereotype specific bullying by building a bullying knowledge base. It is also one of the few articles that discusses end-user strategies such as delayed posting, to prevent this phenomenon in social networks. Zhao et al. [17] created bullying features by applying word embeddings to "insults" and achieved an f-measure of 0.78 with an SVM classifier. In addition to innovative features extracted from a dictionary of common words neurotics use in social networks, Al Garadi et al. [18] used a Twitter dataset and enriched it with social network, user and content features to achieve an f-measure of 0.94 with an SVM classifier.

Deep learning architectures have also started to emerge in cyberbullying detection. Zhang et al. [19] proposed a novel pronunciation based convolutional neural network (PCNN) using phonetic transcriptions as features to minimize the misspelling of words, thus alleviating the problem of bullying data sparsity. Despite of not using word embeddings - which are typically used to represent text in neural networks - the classifier outperforms a baseline CNN (with pre-trained word embeddings) and other algorithms, such as SVM and RandomForests, achieving an f-measure of 0.98 in a Twitter dataset used by Kasture [20] and an f-measure of 0.57 in the Formspring dataset by Reynolds et al. [10] that is used in this paper. Building on Zhang's work, Vishwamitra et al. [21] built a two-level cyberbullying defense mechanism for mobile users, consisting of a pre-send detection scheme that acts as a deterrent to the bully and a fine-grained detection scheme that uses the PCNN to detect subtle forms of cyberbullying attacks.

A very important aspect must be emphasized: for the vast majority of the current state-of-the-art, results are presented as a macro-average between two classes, the Non-Cyberbullying (NCB) category and the Cyberbullying (CB) category. Additionally, most of the presented results are achieved on balanced test datasets. As a consequence, most reported f-measure results are significantly better than they would be if the classifiers were tested to detect Cyberbullying in a real-world scenario, where the number of non-cyberbullying messages is naturally significantly lower than the other types of messages. This data imbalance has been widely documented to affect the predictive capabilities of machine learning classifiers [22] [23]. Nonetheless, this is a highly unbalanced problem, and as such, we argue that: (1) it is utterly useless to report results tested using tailored balanced datasets; (2) the only relevant results are those of the CB category.

III. Fuzzy Fingerprints

In this section, we present what are Fuzzy Fingerprints (FFP) and how they have been adapted to deal with several identification, detection and classification problems.

A. A Brief History

In computer sciences a fingerprint is a procedure that transforms an arbitrarily large data item (e.g. a set of tweets on the same topic) to a compact information block that "uniquely" identifies the original data, the same way human fingerprints can uniquely identify an individual. In text related tasks, the fingerprint of a given class, e.g. the cyberbullying class, is a k-sized ranked vector composed of the most relevant features (typically words) that represent the phenomenon.

TABLE I
EXAMPLE OF A TOP-3 MOST FREQUENT WORDS

Class	Feature	Count
NCB	love	123
	funny	87
	meme	19
CB	kill	142
	hate	59
	die	43

Fuzzy Fingerprints were originally proposed for text classification by Homem et al. [24], where they were successfully used to detect authorship of newspaper articles (out of 73 different authors). In 2014, Rosa et. al. [25] [26] [27] used the fingerprint of several Twitter #hashtags to outperform baseline classifiers in detecting a tweet's topic. Curto et. al. [28] used FFP to improved the prediction of Intensive Care Unit readmissions based on medical text notes and recently, Carvalho et. al. [29] adapted the FFP mechanism to outperform memory-based collaborative filtering Recommender Systems.

B. The Cyberbullying Fuzzy Fingerprint

The base for the implementation of cyberbullying detection using Fuzzy Fingerprints stems from Rosa et. al. [25]. The approach was chosen due to the nature of the data (social networks) and the obtained high recall scores when compared to other approaches, which we hypothesize to be an important metric in the scope of cyberbullying detection.

In order to create a fuzzy fingerprint for a given class/category j, it is necessary to obtain a set of properly classified documents. In the case of our dataset (detailed in Section IV-A), each document/text has two possible classifications: cyberbullying (CB) or no-cyberbullying (NCB). The algorithm computes the top-k word list for each of the 2 classes: all words in the documents are processed to obtain two lists of k tuples $\{v_i, c_i\}$ where v_i is the i-th most frequent word and c_i the corresponding count. i.e., we obtain an ordered k-sized list containing the most frequent distinct words for each category (as an example, please consider Table I excluding stopwords).

In [25], due to the small size of a single tweet and the many possible classes, it is theorized that each word/token in a given class should be as unique as possible, in order to make the fingerprints distinguishable amongst the various classes. Therefore, we calculate the Inverse Class Frequency (icf) of each single word present in all the computed k tuples $\{v_i, c_i\}$. The icf is an adaptation of the well-known Inverse Document Frequency (idf), where classes are used instead of documents to distinguish the occurrence of common words:

$$icf_v = \log \frac{J}{J_v} \tag{1}$$

In Equation (1), J is the size of the fingerprint library (i.e., the total number of different classes), and J_v is the number of classes where word v is present. The product of the frequency of word v with its inverse class frequency, $tficf_v = c_v \times icf$, is used to re-order the k-sized word list

TABLE II
EXAMPLE OF A FINGERPRINT HASH TABLE BEFORE AND AFTER ICF

Class	Feature	Count	Feature	Count × ICF
#refugeeswelcome	refugee	4	refugee	1.90
	migration	2	migration	0.95
	sad	1	sad	0.48
#donaldtrump	president	10	president	4.77
	republican	5	republican	1.43
	trump	1	trump	0.17
#fakenews	truth	8	truth	3.81
	trump	3	media	0.95
	media	2	trump	0.52

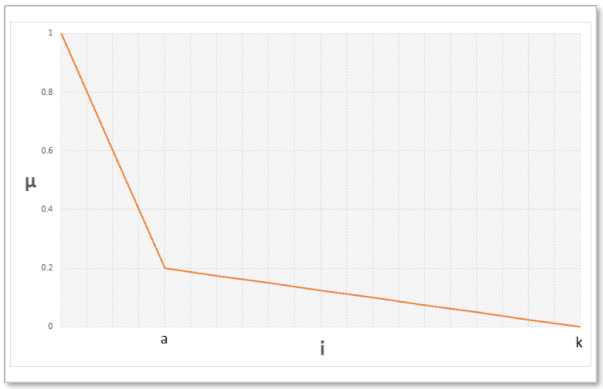

Fig. 1. Fuzzyfying function that determines the membership value μ_{ji} of the i-th ranked word of class j. k is the fingerprint size, and a is 0.2*k

of each topic. Table II shows an example of the application of the icf when considering 3 classes (3 different hashtags) and fingerprint size $k = 3$. The example shows a possible top-k output produced by the algorithm after going through a small training set. By multiplying the occurrences of each word per topic with its icf, we obtain the third column of Table II. As expected, the term "trump" drops one position in the ranking of words for the topic "#fakenews", since it was the only word occurring in more than one fingerprint. When applied to the current cyberbullying detection problem, where only 2 classes are possible, the ICF demotes words that are common to both fingerprints (i.e. moves them further down each top-k list).

The next step to obtain the FFP consists in fuzzifying each top-k list. In [25], a fuzzifying function inspired by the Pareto rule was used (Equation (2) - Figure 1). Roughly 80% of the membership value is assigned to first 20% elements in the ranking, and the remaining 80% are assigned less than 20% of the membership value. In Equation (2), μ_{ji} is the membership value of the i-th ranked word of class j, k is the fingerprint size, $i = [0, ..., k]$ and $a = 0.2 * k$. For example, for $k = 10$ and $a = 2$.

$$\mu_{ji} = \begin{cases} 1 - \frac{(1-\frac{a}{k}) \times i}{a} & i < a \\ \frac{a(1-\frac{i-a}{k-a})}{k} & i \geq a \end{cases} \tag{2}$$

The fingerprint of class j, Φ_j is based on the top-k list, and consists of a size-k fuzzy vector where each position contains an element v_{ji} (a word of class j), and a membership value μ_{ji} representing the fuzzified value of the rank of v_{ji} (the

membership of the rank), obtained by the application of (2). Formally, class j will be represented by its size k fingerprint $\Phi_j = \{(v_{j1}, \mu_{j1}), (v_{j2}, \mu_{j2}), ..., (v_{jk}, \mu_{jk})\}$. For the purpose of this work, two fingerprints will constitute the fingerprint library: the cyberbullying fingerprint and the no-cyberbullying fingerprint.

The classification of a given text is based on the similarity of the text to the fingerprint of each class. The similarity function is shown in (3), where Φ_j is the fingerprint of class j, T is the set of distinct words of the preprocessed text, $S_{\Phi_j} = \{v_{j1}, v_{j2}, ..., v_{jk}\}$ is the set of words of fingerprint Φ_j, and μ_{ji} is the membership degree of word v_{ji} in the fingerprint Φ_j. Essentially, T2S2 consists of adding the fuzzy membership values of every word v that is common between the document and the fingerprint Φ_j, and then normalizing the result by dividing it with the sum of the top x membership values of fingerprint Φ_j (x is the minimum between k and the cardinality of T). T2S2 tends to 1 when most to all features of the tweet belong to the top words of the fingerprint, and approaches 0 when there are no common words between the tweet and the fingerprint, or the few common words are in the bottom of the fingerprint.

$$\mathrm{T2S2}(T, \Phi_j) = \frac{\sum\limits_{v_{ji}} \mu_{ji} : v_{ji} \in (T \cap S_{\Phi_j})}{\sum\limits_{i=0}^{\min(\#T,k)} \mu_{ji}} \tag{3}$$

Since only two fingerprints were created for the cyberbullying detection approach, we altered the method used in [25] to classify and retrieve relevant information. In this work we compute the similarity to each of the classes CB and NCB, and select the higher one to indicate the most probable class. However the text is only retrieved from the database as a CB case if the T2S2 similarity is above a given threshold. This threshold and other FFP relevant parameters are discussed in the following section. Please consider this very simple example with cyberbullying class from Table I, after being fuzzyfied (Table III).

TABLE III
FUZZYFYIED TOP-3 CYBERBULLYING FINGERPRINT

Feature	μ_{ji}
kill	1.0
hate	0.1667
die	0.083

Calculating the T2S2 similarity score of the obviously aggressive sentence *"I will kill your mother"*, which after stopword removal would contain only 2 features {kill, mother}, would yield (4), with a very high similarity score to the CB class.

$$\mathrm{T2S2} = \frac{1.0}{1.0 + 0.1667} = 0.857 \tag{4}$$

IV. EXPERIMENTAL SETUP

A. Dataset

This work uses an extended version of the Formspring dataset from [10], available at Kaggle[1].

Formspring[2] is a question-and-answer-based social network launched in 2009 that was the genesis for other similar websites, such as AskFM[3] and Tumblr[4]. The data available represents 50 ids that were crawled in Summer 2010. For each id, the profile information and each post (question and answer) was extracted. Each post was then loaded into Amazon's mechanical turk and labeled by three workers for cyberbullying content, in a total of 13160 labeled texts. Out of this total, 2205 texts were deemed by at least one worker to contain cyberbullying, while 10955 show no evidence of said phenomenon. The complete vocabulary contains 17846 tokens consisting of words, emoticons, etc.

Each train and test sample is a full string containing both the question and the answer in the same sequence, with the question (Q:) and answer (A:) portion of it, perfectly identified: *"Q: Are you a morning or night person?
A: Night 4shuree!!"*. We preprocess the text before any further transformation. Namely:

- the *"Q:"* and *"A:"* markers are removed;
- any html tag or encoding representation are removed (example: *
*);
- words with characters that repeat more than twice are normalized to keep at most two repetitions (example: *goooood* becomes *good*);

Three versions of the dataset are used in the experiments in order to show how relevant for this task is the balancing of data:

- a down-sampled balanced version consisting of all the CB class entries (2205) and a random selection of the NCB class with the same size, for a total of 4410 samples (vocabulary length of 8766 items);
- a full version with the unbalanced class ratio as described above;
- a balanced training dataset consisting of 3696 texts with a unbalanced testing dataset with 205 cyberbullying samples and 1110 no-cyberbullying samples;

Despite not fully explored in this work, there are solutions to handling the reported label imbalance. The simplest option is to downsample the majority class (as explained above), but one could also generate synthetic samples in an attempt to upsample the minority class or study the use of penalized learning algorithms that increase the cost of classification mistakes on the minority class.

[1]https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection
[2]http://formspring.me
[3]https://ask.fm
[4]https://www.tumblr.com

B. Parameter Optimization

All of the experiments performed below are subjected to a 10-fold cross-validation. In each fold, a grid search for best performing parameter combination is executed on the training set. The following are Fuzzy Fingerprints' parameters eligible for optimization, and their possible values:

- size k of the fingerprint: {20, 100, 500, 1000, 2500, 5000}
- minimum word length: {1, 2, 3}
- threshold of the T2S2: {0.1, 0.25, 0.5}
- removal of stopwords: {YES; NO}
- calculate the Inverse Class Frequency (ICF): {YES; NO (i.e., word count per class)}

For benchmarking purposes, the experiment also includes an SVM, Multinomial Naive Bayes and Logistic Regression classifier tests, with samples represented as a sparse tf-idf vector bag-of-words and the following parameter grid search deployed: tokens {unigrams, unigrams + bigrams, bigrams}, stopword removal {yes, no} and the regularization trade-off parameter C {0.001, 0.01, 0.1, 1.0, 10.0} (for SVM and Logistic Regression) or the alpha smoothing parameter α {0.1, 0.25, 0.5, 0.75, 1.0} for MNB.

V. Results and Analysis

A. Balanced Dataset

In this section we present the obtained results when approaching the problem as a typical binary classification task, which is the norm in related works, as explained in the Introduction. We use a balanced dataset, with all classifiers trained and tested with an equal number of samples from each category. We emphasize that this is not a real world scenario, and any conclusions are irrelevant for a real world application of cyberbullying detection. The use of a balanced testing scenario for this experiment, was made purely out of academic interest to exemplify the high values of Precision and Recall shown in the literature.

The best parameter configuration used for the FFP in all folds was as follows: Min. word length= 1; $k = 2500$; threshold= 0.10; no ICF and no stopword removal - 10/10 folds

A high value of k is needed, a low threshold value is best and no ICF or stopword removal is performed. Additionally, the minimum word length is 1, i.e., no words are removed based solely on their length, which means the whole vocabulary is used. We theorize that this is a an effort of the algorithm to include words in the fingerprints that match those of the text being evaluated. Because a shorter dataset (4405 samples versus 13160 samples) will provide a smaller and poorer vocabulary, it is more likely that test samples may appear that have no match with any of the fingerprints. In this instance, by including the totality of the vocabulary, it increases the odds of the FFP having enough (or more important) keywords to make a prediction.

The results for the experiment are presented in Table IV.

TABLE IV
Method Comparison - Balanced Dataset

Method	Precision	Recall	F-measure
Fuzzy Fingerprints	0.811	0.810	0.810
Logistic Regression	0.826	0.823	0.823
Multinomial Naive Bayes	0.817	0.816	0.815
SVM	0.824	0.823	0.823

It is possible to see that all classifiers achieve similar and rather good results. Differences are small in both Precision and Recall (and consequently F-measure) with no discernible differences allowing to conclude that a given baseline classifier is better than the others.

B. Unbalanced Dataset

In this section we use the fully unbalanced version of the dataset (both train and test sets are unbalanced), that technically approaches a more realistic scenario.

Parameter optimization allowed for some interesting conclusions, since while for SVM and MNB, the parameter grid search gives the exact same combination regardless of the fold, for Logistic Regression and FFP, there are variations:

- SVM - $C = 10.0$, tokens = unigrams + bigrams and stopword removal
- Logistic Regression:
  1) $C = 10.0$, tokens = unigrams + bigrams, stopword removal - 6 out of 10 folds
  2) $C = 10.0$, tokens = unigrams, no stopword removal - 4 out of 10 folds
- MNB - $\alpha = 0.1$, tokens = unigrams and stopword removal
- FFP:
  1) Min. word length= 3, $k = 5000$, threshold= 0.10, no ICF and no stopword removal - 5 out of 10 folds
  2) Min. word length= 3, $k = 2500$, threshold= 0.10, no ICF and no stopword removal - 2 out of 10 folds
  3) Min. word length= 3, $k = 5000$, threshold= 0.25, no ICF and no stopword removal - 2 out of 10 folds
  4) Min. word length= 3, $k = 5000$, threshold= 0.10, no ICF and stopword removal - 1 out of 10 folds

The FFP parameter optimization performed in each fold shows some common trends with the tests performed in the balanced dataset from the previous section: much like in previous works [25] [26] [27] removing stopwords is detrimental to performance, only word frequency should be used to order the fingerprint features and words must have at least 3 characters. This means that a word such as "the" will probably be ranked high in both the fingerprint for the CB class and the NCB class. This phenomenon is also valid to explain the high $k = 5000$ value, in contrast to the reported $k = 20$ by [27]. Because such words that have very little distinctive influence may occur in both fingerprints, a higher k will ultimately force the algorithm to include them, as well as others with more relevance to the concept of cyberbullying. Curto et. al. [28] had already documented the need for a higher k-sized fingerprint

when dealing with a binary classification problem via FFP. In contrast, a small k value would be insufficient to represent the key words in either class, both because of the complexity of the cyberbullying phenomenon and the broadness of the no-cyberbulling category.

In what concerns the obtained results, we start by presenting the commonly used "overall performance", consisting of the macro-average between the two classes CB and NCB (Table V).

These "overall results" seem to largely penalize the use of the FFP to address this problem: with the exception of the FFP, all competing methods achieve similar results. Furthermore, the results are improved over the balanced dataset (except for the Fuzzy Fingerprints, which are indeed much worse)!

TABLE V
METHOD COMPARISON - UNBALANCED DATASET - MACRO AVERAGE OF CB AND NCB CLASSES

Method	Precision	Recall	F-measure
Fuzzy Fingerprints	0.829	0.750	0.773
Logistic Regression	0.840	0.861	0.838
Multinomial Naive Bayes	0.822	0.842	0.816
SVM	0.834	0.850	0.838

However, these apparently good results are a consequence of the unbalanced nature of the dataset. The performance in classifying the majority category (NCB) is unnecessary and creates a heavy bias, not only because a classifier can become more proficient in predicting a certain class if it has more samples of said class to train from, but also because the reported metrics are macro-metrics weighted by the label imbalance of the dataset (as mentioned in section IV-A, the number of no-cyberbullying samples is approximately 5 times larger).

TABLE VI
UNBALANCED DATASET - CYBERBULLYING CLASS

Method	Precision	Recall	F-measure
Fuzzy Fingerprints	**0.355**	**0.597**	**0.425**
Logistic Regression	0.594	0.318	0.394
Multinomial Naive Bayes	0.538	0.243	0.308
SVM	0.525	0.377	0.421

If one looks exclusively at the CB class (Table VI), which is the only one relevant for the "Cyberbullying detection" problem, the results are totally different and differ significantly among the several techniques. FFP is now the best technique (despite only 1.0% better in absolute f-measure), closely followed by SVM, with the other techniques lagging behind.

A very important difference between the best methods (FFP and SVM) lies in how the f-measure value is achieved. FFP rely mostly in a good recall value, while SVM (and the others) rely mostly in a good precision. This is relevant because it is more important to have a good recall than a good precision when detecting cyberbullying: recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances, indicating how good a method is at retrieving the cyberbullying cases (our goal), while precision indicates the percentage of cyberbullying cases among all retrieved instances. Note that even though we would like to ideally detect all cyberbullying cases, this cannot be done at the cost of precision with the risk of manually analyzing millions of irrelevant messages in a real world scenario. Hence, even though the recall is more important, as Reynolds et. al. [10] also suggests, there is a need to maximize the f-measure (and not the recall). The use of the threshold mechanics applied to the T2S2 score can explain the FFP behavior in what concerns the better recall: a low threshold implies a more relaxed detection of cyberbullying in texts with fewer relevant words, which in turn may increase the number of False Positives (lowering precision) and decrease the number of False Negatives (increasing recall).

Independently from the fact that FFP performs slightly better, it should be noted that the detection of cyberbullying occurrences based only on these simple features (single messages) is rather poor in this more realistic unbalanced scenario: f-measure results drop from 0.8 to around 0.4, which is not acceptable for a real-life cyberbullying detection in social networks application.

C. Balanced Train and Unbalanced Test

Because further argument could be made that a balanced training dataset could help distinguishing between two classes while testing performance in an unbalanced dataset, a third a final experiment is detailed in this section.

TABLE VII
TRAIN: BALANCED - TEST: UNBALANCED - CB CLASS

Method	Precision	Recall	F1-Score
Fuzzy Fingerprints	0.274	0.731	0.390
Logistic Regression	0.260	0.751	0.379
MNB	0.281	0.727	0.396
SVM	0.255	0.760	0.379

Table VII shows that training the classifiers with balanced data results in a large increase in recall for all tested techniques (which is good), but at an high cost of precision resulting in a lower f-measure compared to the fully unbalanced scenario (Table VI).

VI. CONCLUSIONS AND FUTURE WORK

In this work we have shown how two key ideas regarding cyberbullying detection affect the interpretation of the achieved results and the implementation of a possible real world detection system: (i) cyberbullying detection is not a binary task, i.e., only the results relating to the CB class are needed to evaluate a cyberbullying detection system; (ii) regardless of the balancing of data in the training set, the test must always be performed in a nearest to real life case study as possible, i.e., with heavy unbalancing reflected in much fewer CB test samples. Based on these ideas we have shown that most of the previously published ML works concerning automatic detection of cyberbullying in social networks present highly biased results that are not suitable for a real world application.

Based on these ideas, an in order to improve the baseline approach for cyberbullying detection, we tested how Fuzzy Fingerprints (FFP) behave when addressing this problem, and concluded that FFP are able to beat standard ML baselines: the FFP achieve an f-measure $= 0.425$ is approximately 1% better than the second best performing classifier (SVM), which while might seem too little of an improvement, is improvement nonetheless. In addition, the FFP achieve these results with a much higher recall, which is an important fact to consider.

Despite the above results, we recognize that, contrary to most of the what related work state, the problem of automatic cyberbullying detection in real world social networks is far from being solved, since the f-measure for the cyberbullying class (the only relevant one) is still largely under 0.5 even for the best classifiers using expanded features. Additionally, we hypothesize that cyberbullying detection cannot rely solely in a single exchanged message, and is highly dependent on previous context and interactions. As a future work, we intend to adapt FFP to a dataset that better captures: (i) the notion of repetition in cyberbullying (as opposed to isolated agression); (ii) context between the people involved in an online conversation (we've found that friends may be aggressive to each other, but as long as both parties take it as a joke, it is OK).

ACKNOWLEDGMENTS

This work was supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project PTDC/MHCPED/3297/2014, project PTDC/IVC-ESCT/4919/2012 and project UID/CEC/ 50021/2013.

REFERENCES

[1] D. Richardson and C. F. Hiu, *Ending the torment: tackling bullying from the schoolyard to cyberspace*. 2016.

[2] B. Belsey, "Cyberbullying: An emerging threat to the always on generation," *Bullying Org. Canada*, pp. 1–9, 2005.

[3] J. Amado, A. Matos, and T. Pessoa, "Cyberbullying: um desafio à investigação e à formação," *Revista Interacções*, vol. 13, no. 13, pp. 301–326, 2009.

[4] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," *Computers in Human Behavior*, vol. 29, no. 1, pp. 26–32, 2013.

[5] R. Slonje and P. K. Smith, "Cyberbullying: Another main type of bullying?: Personality and Social Sciences," *Scandinavian Journal of Psychology*, vol. 49, no. 2, pp. 147–154, 2008.

[6] Q. Li, "A cross-cultural comparison of adolescents' experience related to cyberbullying," *Educational Research*, vol. 50, no. 3, pp. 223–234, 2008.

[7] H. Jang, J. Song, and R. Kim, "Does the offline bully-victimization influence cyberbullying behavior among youths? Application of General Strain Theory," *Computers in Human Behavior*, vol. 31, no. 1, pp. 85–93, 2014.

[8] P. C. Ferreira, A. M. V. Sima??, A. Ferreira, S. Souza, and S. Francisco, "Student bystander behavior and cultural issues in cyberbullying: When actions speak louder than words," *Computers in Human Behavior*, vol. 60, pp. 301–311, 2016.

[9] J. Bayzick, A. Kontostathis, and L. Edwards, "Detecting the Presence of Cyberbullying Using Computer Software," *Springer*, no. December, pp. 11–12, 2011.

[10] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, vol. 2, pp. 241–244, 2011.

[11] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," *Association for the Advancement of Artificial Intelligence*, pp. 11–17, 2011.

[12] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving Cyberbullying Detection with User Context," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7814 LNCS, pp. 693–696, 2013.

[13] V. S. Chavan and S. S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," *2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015*, pp. 2354–2358, 2015.

[14] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber Bullying Detection Using Social and Textual Analysis," *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia - SAM '14*, pp. 3–6, 2014.

[15] V. Nahar, X. Li, and C. Pang, "An Effective Approach for Cyberbullying Detection," *Communications in Information Science and Management Engineering*, vol. 3, no. 5, pp. 238–247, 2014.

[16] K. Dinakar, R. Picard, and H. Lieberman, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2015-Janua, no. 3, pp. 4168–4172, 2015.

[17] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," *Proceedings of the 17th International Conference on Distributed Computing and Networking - ICDCN '16*, pp. 1–6, 2016.

[18] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.

[19] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon, "Cyberbullying detection with a pronunciation based convolutional neural network," *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, no. February 2017, pp. 740–745, 2017.

[20] A. Kasture, *A predictive model to detect online cyberbullying*. PhD thesis, Auckland University of Technology, 2015.

[21] N. Vishwamitra, X. Zhang, J. Tong, H. Hu, F. Luo, R. Kowalski, and J. Mazer, "MCDefender: Toward effective cyberbullying defense in mobile online social networks," *IWSPA 2017 - Proceedings of the 3rd ACM International Workshop on Security and Privacy Analytics, co-located with CODASPY 2017*, pp. 37–42, 2017.

[22] N. V. Chawla, N. Japkowicz, and P. Drive, "Editorial : Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[23] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook*, pp. 875–886, Boston, MA: Springer US, 2009.

[24] N. Homem and J. P. Carvalho, "Authorship identification and author fuzzy "fingerprints"," *Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS*, pp. 180–185, 2011.

[25] H. Rosa, F. Batista, and J. P. Carvalho, "Twitter Topic Fuzzy Fingerprints," in *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 776–783, IEEE, jul 2014.

[26] H. Rosa, J. Carvalho, and F. Batista, "Detecting a tweet's topic within a large number of portuguese twitter trends," in *OpenAccess Series in Informatics*, vol. 38, 2014.

[27] H. Rosa, *Topic Detection within Public Social Networks*. Master thesis, Instituto Superior Técnico - Universidade de Lisboa, 2014.

[28] S. Curto, J. P. Carvalho, C. Salgado, S. M. Vieira, and J. M. C. Sousa, "Predicting ICU readmissions based on bedside medical text notes," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 2144–a–2151–h, IEEE, jul 2016.

[29] A. Carvalho, P. Calado, and J. P. Carvalho, "Combining ratings and item descriptions in recommendation systems using fuzzy fingerprints," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6, IEEE, jul 2017.