Fuzzy Multi-task Learning for Hate Speech Type Identification

Han Liu*
School of Computer Science and Informatics, Cardiff University
Cardiff, United Kingdom
liuh48@cardiff.ac.uk

Pete Burnap
School of Computer Science and Informatics, Cardiff University
Cardiff, United Kingdom
burnapp@cardiff.ac.uk

Wafa Alorainy
School of Computer Science and Informatics, Cardiff University
Cardiff, United Kingdom
alorainyws@cardiff.ac.uk

Matthew L. Williams
School of Social Science, Cardiff University
Cardiff, United Kingdom
williamsm7@cardiff.ac.uk

ABSTRACT

In traditional machine learning, classifiers training is typically undertaken in the setting of single-task learning, so the trained classifier can discriminate between different classes. However, this must be based on the assumption that different classes are mutually exclusive. In real applications, the above assumption does not always hold. For example, the same book may belong to multiple subjects. From this point of view, researchers were motivated to formulate multi-label learning problems. In this context, each instance can be assigned multiple labels but the classifiers training is still typically undertaken in the setting of single-task learning. When probabilistic approaches are adopted for classifiers training, multi-task learning can be enabled through transformation of a multi-labelled data set into several binary data sets. The above data transformation could usually result in the class imbalance issue. Without the above data transformation, multi-labelling of data results in an exponential increase of the number of classes, leading to fewer instances for each class and a higher difficulty for identifying each class. In addition, multi-labelling of data is very time consuming and expensive in some application areas, such as hate speech detection. In this paper, we introduce a novel formulation of the hate speech type identification problem in the setting of multi-task learning through our proposed fuzzy ensemble approach. In this setting, single-labelled data can be used for semi-supervised multi-label learning and two new metrics (detection rate and irrelevance rate) are thus proposed to measure more effectively the performance for this kind of learning tasks. We report an experimental study on identification of four types of hate speech, namely: religion, race, disability and sexual orientation. The experimental results show that our proposed fuzzy ensemble approach outperforms other popular probabilistic approaches, with an overall detection rate of 0.93.

*The corresponding author

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313546

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Sociology**.

KEYWORDS

Machine learning; Multi-task learning; Cyberhate detection; Text classification; Fuzzy classification

ACM Reference Format:
Han Liu, Pete Burnap, Wafa Alorainy, and Matthew L. Williams. 2019. Fuzzy Multi-task Learning for Hate Speech Type Identification. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313546

1 INTRODUCTION

Traditional supervised learning typically involves training of a classifier on a single-labelled data set, i.e. each instance is assigned one label only. In this context, each label is treated as a single class and different classes are assumed to be mutually exclusive. However, this assumption does not always hold in real world applications. For example, images can be labelled to show different concepts, different people and different objects. Also, a movie can be included in multiple categories. Therefore, researchers have been motivated to transform the problem to multi-label classification (learning).

In the context of multi-label learning, each instance can be assigned more than one label. Each distinct set of labels *LS* assigned to one or more instances can be generally treated as a single class C_k, such that the training of a classifier can still be undertaken through discriminating one class from the other classes. This strategy is referred to as 'Label Power-set' [1] in the setting of single-task learning, which aims to allow a traditional learning approach to be used for training a single classifier on multi-labelled data [21] as a single task. On the other hand, a multi-labelled data set D can also be transformed into n binary data sets $D_1, D_2, ..., D_n$ (each one D_k per label l_k), such that a traditional learning approach can be used to train a binary classifier h_k on each data set D_k for identifying the corresponding label l_k. This strategy is referred to as 'Binary Relevance' [39] in the setting of multi-task learning and the training of the binary classifiers is treated as multiple tasks [21]. A more detailed review of multi-label learning can be found in [32, 39].

In general, different labels can have specific relationships such as mutual independence and positive correlation (co-occurrence) , if the labels are not mutually exclusive. In this context, if the labels assigned to each instance are mostly independent of each other, the above Label Power-set strategy for handling multi-labelled data could result in an exponential increase of the number of classes and a reduced number of instances for each class [22, 32]. In this case, the complexity of the learning task is much increased leading to a higher risk of overfitting [29]. Also, if the labels assigned to each instance have some positive correlations, the above Binary Relevance strategy through data transformation would not only result in the class imbalance issue but also fail to identify potential correlations between labels [21, 29]. In addition, data labelling is very time consuming and expensive in some application areas such as hate speech detection [2], which indicates a higher degree of infeasibility of multi-labelling of data in these areas.

In order to overcome the above limitations, in this paper, we propose a three-level framework for hate speech detection, which involves identification of the presence or absence of hate speech in level 1, identification of the types of hate speech in level 2 and identification of topics and contexts of hate speech in level 3. However, we focus on level 2, since different types of hate speech could have some intersectionality [9] increasing the complexity of this problem and an appropriate formulation of this problem is needed to achieve effective identification. In particular, we propose a fuzzy ensemble approach for hate speech type identification in the setting of multi-task learning. In this setting, we show how single-labelled data can be used to enable semi-supervised multi-label learning through a new problem formulation and two new metrics are also defined for effective evaluation of the performance for this kind of learning tasks. An experimental study is reported to show that the proposed fuzzy approach is more suitable and effective for this kind of learning tasks, comparing with the popular approaches.

The rest of this paper is organized as follows: Section 2 provides an overview of cyberhate research. In Section 3, we introduce the proposed three-level framework in general and focus on illustration and justification of the proposed fuzzy approach for semi-supervised multi-label learning. In Section 4, we provide the general description of the data used for this experimental study and show the results for evaluation of the proposed approach in comparison with the state of the art ones. In Section 5, the contributions of this paper are summarized and further directions are suggested towards further advances.

2 RELATED WORK

Since the spread of online hate speech could lead to disruptive anti-social outcomes, cyberhate has thus been considered as a legal issue in many countries [5]. In particular, many European countries have already initiated legal actions to prevent online hate speech to be posted [6]. However, it has been complicated to take such actions since the World Wide Web is naturally borderless [24]. Also, due to different laws from different countries, it has become more difficult to prosecute the senders of online hate speech and this even results in the lack of power for removing any hateful contents posted from a location outside their territory [5]. Outside of legal procedures, social network providers, such as Facebook and Twitter, have also been responsible to take any necessary steps towards restricting online hate speech significantly [4]. This has motivated researchers to develop automatic tools for hate speech detection, such that the post of hateful contents can be effectively prevented [2]. In particular, machine learning has become a very popular tool for automatic detection of hate speech [8, 35, 40].

In the context of machine learning based cyberhate detection, some traditional learning algorithms, such as Support Vector Machines (SVM) [9], Naive Bayes (NB) [14, 28], Logistic Regression (LR) [33, 35] and Random Forests (RF) [8], have also been used in the previous studies. A pragmatic approach was proposed in [34] for detecting hateful and offensive expressions, based on unigrams and automatically collected patterns for training classifiers by using SVM, DT and RF. Also, some other state-of-the-art methods of feature extraction, such as N-grams (NG) [31] and Typed Dependencies (TD) [8, 9] and Text Embedding [25], have been used to capture hateful characteristics. In general, text embedding shows its advantages (in terms of reduction of the dimensionaltiy and the sparsity) and better performance than the other methods, while the three learning algorithms SVM, NB and DT generally show their better suitability for training classifiers.

In recent years, deep learning methods have also been used for both feature extraction and training of classifiers. In particular, Gamback and Sikdar have recently used Convolutional Neural Networks (CNN) in [12] for classifying hate speech using different types of features and they showed that the use of embedding features extracted through word2vec led to the best performance on a data set that involves multiple classes [33]. In [3], multiple deep neural network (DNN) architectures, such as CNN and Long Short Term Memory (LSTM) Networks, were compared using the same data set. The DNN architectures were adopted to learn semantic word embeddings as features for training classifiers. The experimental results reported in [3] show that the use of embedding features led to better classification performance than the use of features extracted through BOW or NG. In terms of training classifiers on the above embedding features, the results show that GBT outperformed DNNs and traditional learning methods (SVM and LR).

A two-step CNN based hybrid approach was proposed in [26] for detecting racist and sexist speech using the same data set as [33]. In particular, the first step is aimed at detection of abusive language, whereas the second step is aimed at identification of the abusive language type (racist and sexist). In comparison of the two-step approach with the one step approach, the experimental results reported in [26] show that the use of hybrid CNN led to the best performance through the one-step approach and the use of LR led to the best performance through the two-step approach, where the one-step approach performed marginally better than the two-step approach. In [40], a gated recurrent unit layer was incorporated into CNN by locating this added layer after the pooling layer and a drop out layer was used for optimaizing the training of embeddings. The experimental results show that the proposed approach led to advances in the classification performance on 6 out of 7 data sets, in comparison with previous baseline results.

On the basis of the above review, we consider text embedding as the state of the art method of feature extraction. Moreover, probabilistic learning approaches including DT, NB, SVM, GBT and DNNs are considered as the state of the art ones for classifiers training.

3 PROPOSED FRAMEWORK OF CYBERHATE DETECTION

In this section, we present the proposed three-level framework for cyberhate detection, which involves polarity classification (hate or non-hate) in level 1, multi-task classification for identification of the hate speech types in level 2 and detection of topics and contexts of hate speech in level 3. Then we focus on level 2 for the problem formulation and the illustration of the procedure of the proposed fuzzy ensemble approach in the setting of multi-task learning.

The proposed framework essentially aims at different levels of abstraction for hate speech detection tasks and reduction of the complexity of a single task in a single level. There has been a plenty of works done on general detection of hate speech as a task involved in level 1, as reviewed in Section 2. However, to the best of our knowledge, very few works have been focused on identifying the presence or absence of multiple types of hate speech as a task involved in level 2. Different types of hate speech have potential intersectionality [9], which indicates that the hate speech type identification task can not be simply formulated as a traditional multi-class classification problem based on the assumption of mutual exclusion of different classes. In other words, multi-label classification would be a more appropriate formulation.

Multi-label classification is typically achieved through training classifiers in the setting of supervised learning. However, mutlilabelling of hate speech data is practically much less feasible, as stressed in Section 1. Therefore, we will introduce a new formulation of the problem in Section 3.1 to enable multi-label classification in the setting of semi-supervised learning and propose a fuzzy ensemble approach for hate speech type identification in the setting of mutli-task learning in Section 3.2.

3.1 Problem Formulation

In general, a single-labelled data set D involving n labels could be produced by taking the positive instances from n different binary data sets $D_1, D_2, ..., D_n$. In this way, each set of positive instances is associated with one of the n labels and the n sets of positive instances are merged into a new single-labelled data set D.

In the setting of semi-supervised multi-label learning, the problem is formulated as n Positive-Unlabelled (PU) learning sub-problems (Eqs. (1) and (2)), respectively, for identifying n labels (types of hate speech), similar to the Binary Relevance based problem transformation. In particular, for each single-labelled instance, it is considered as an unknown case that the instance is relevant or not to the other $n-1$ labels, i.e. it is treated as unlabelled regarding other labels.

$$h(\cdot) = \bigwedge_{j=1}^{n} \{h_j(\cdot)\} \tag{1}$$

$$h_j(\cdot) = \begin{cases} +1, & \text{if } h_j(\cdot) = l_j; \\ -1, & \text{otherwise.} \end{cases} \tag{2}$$

In the above context, for each predefined label $l_j \in LS$, instances that are assigned this label l_j are considered to be a set of positive ones, whereas the other instances are considered to be a set of unlabelled ones regarding the label l_j (i.e. a mixed set of potentially both positive and negative instances). The above two sets of instances are used for an independent binary classification task for each label l_j in the setting of PU learning [13, 15–17], which has been popularly adopted for semi-supervised single-label classification.

PU learning is a special type of semi-supervised learning, which does not involve labelled negative instances in the training set and can be achieved through three strategies as surveyed in [38]:

(1) The first strategy involves a two-step operation by identifying reliable negative instances from the unlabelled set and then adopting supervised learning on the positive instances and the reliable negative instances.
(2) The second strategy involves weighting the positive and unlabelled instances and then estimating the conditional probability of the positive class given the input vector of an instance.
(3) The third strategy involves simply treating the unlabelled training instances as highly noisy negative ones, assuming that the majority of the unlabelled instances belong to the negative class.

In the context of hate speech type identification, we take the third strategy since the positive (target) class is typically the minority one. For each binary classifier h_j trained in the setting of PU learning, the output +1 shown in Eq. (2) indicates the case that the target type l_j of hate speech is detected from an instance x_i. In contrast, the other output -1 indicates the case that the binary classifier h_j fails to detect the target type l_j of hate speech.

For adopting probabilistic approaches, it is not feasible to achieve the above setting of semi-supervised multi-label learning without transformation of the original single-labelled data. In particular, a single-labelled data set needs to be transformed into n binary data sets in the setting of Binary Relevance, i.e. from each binary data set D_j, the instances assigned label l_j are treated to be the positive ones and the other instances are treated as unlabelled ones, so a binary classifier h_j is trained to identify the label l_j.

In contrast, for adopting the proposed fuzzy approach, the above data transformation is not needed, since it is an instance based generative approach of fuzzy rule learning, which aims at considering iteratively whether each instance $x_i \in l_j$ is covered by an existing rule of the label l_j or another label $l_g \neq l_j$. More details on the fuzzy approach are shortly given in Section 3.2.

In the classification stage, it is possible that more than one binary classifier gives a positive output (i.e. a label $l_j \in LS$ rather than its negation $\neg l_j$) for the same instance x_i, so the instance is finally assigned multiple labels. If the target label l_t is originally assigned to the instance x_i as the ground truth and the binary classifier h_t successfully identifies that the instance x_i belongs to the label l_t, then it would be judged as successful detection of the target type l_t of hate speech from the instance x_i, regardless of the correctness of assigning other labels to the instance x_i. Also, it is possible that none of the binary classifiers provides a positive output, which would be simply identified as the case that the binary classifier h_t fails to detect the presence of the target type l_t of hate speech and none of the other types of hate speech is detected.

In the new problem formulation shown in Eqs. (1) and (2), two new metrics are proposed, which are referred to as 'detection rate' (Eq. (3)) and 'irrelevance rate' (Eq. (4)), respectively. The two terms given for the two metrics are inspired from two possible cases in information retrieval tasks, where the retrieved documents may be the target ones successfully detected or the irrelevant ones incorrectly identified. In Eq. (3), $|l_j|$ represents the number of instances

that are annotated the label l_j and $|l_j \cap \hat{l_j}|$ is the number of instances that are annotated the label l_j by people and are assigned either a single label l_j or multiple labels including l_j by classifiers. In Eq. (4), $|\hat{l_j}|$ is the number of instances that are assigned the label l_j by classifier h_j and $|\neg l_j|$ represents the number of instances that are annotated as $\neg l_j$ (i.e. the number of negative instances).

$$P = \frac{|l_j \cap \hat{l_j}|}{|l_j|} \tag{3}$$

$$R = \frac{|\hat{l_j}|}{|\neg l_j|} \tag{4}$$

In the context of hate speech types identification, non-hate speech instances should be used as negative ones (belonging to $\neg l_t$), since single-labelled hate speech instances (annotated label l_t) are not provided with ground truth on the absence of each other type $l_j \neq l_t$ of hate speech and thus cannot be used as negative instances for evaluating the overall confidence of relevantly assigning an instance $x_i \in l_t$ a label $l_j \neq l_t$ in addition to the label l_t.

3.2 Procedure of Multi-task Learning

The proposed multi-task learning approach is essentially based on the mixed fuzzy rule formation algorithm [7] with modifications, and the procedure of this algorithm is illustrated as below:

It involves a sequential and constructive generation of new rules and modification of existing rules in an instance-by-instance manner, i.e. each instance is checked, and a new rule is added into the rule set or some existing rules are modified. For each class label l_j, a subset of fuzzy rules is trained in a single task and the rule subset is treated as a binary classifier h_j trained for identifying the label l_j. Therefore, the training of the whole set of fuzzy rules for all class labels is undertaken in a multi-task learning manner.

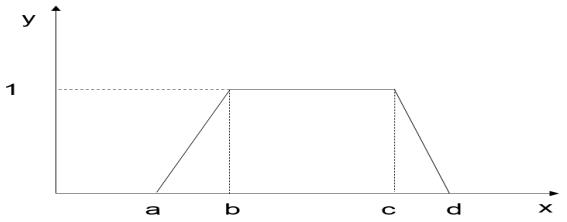

Figure 1: Trapezoidal Membership Function [19]

In the whole procedure, each rule r_t involves n membership functions (for n rule antecedents) and two additional parameters w and λ to be defined, where w represents the number of instances covered by rule r_t and λ is a so-called anchor that remembers the original instance triggering the generation of this rule r_t. A membership function generally involves four parameters a, b, c, d as shown in Fig. 1, where the interval $[b, c]$ represents a core region showing hard boundaries for an element to fully belong to a set and the rest represents a support region (a, b) or (c, d) showing soft boundaries for an element to partially belong to a set.

At each of a number of epochs, once each instance $x_i \in l_j$ is checked, one of the following cases needs to be identified:

- Covered: if an instance $x_i \in l_j$ is covered by a rule r_t, then the membership function defined for each rule antecedent needs to be adjusted to let the instance x_i obtain a membership degree of 1 to the rule r_t [7, 11].
- Commit: if the instance $x_i \in l_j$ is not covered by any rules, a new rule is generated and a membership function is initialized for each rule antecedent [7, 11].
- Shrink: if an instance $x_i \in l_j$ is covered by a rule r_t of label $l_g \neq l_j$, then conflict avoidance should be taken using some shrink heuristics [11].

Following the above procedure, a set of fuzzy rules is trained, which involves n subsets of rules for the n labels. The presence or absence of each target type of hate speech can be identified from each new instance through fuzzification, inference and defuzzification. The fuzzification operation is simply to map the value v_{ik} of each feature f_k of an instance x_i to a membership degree $\mu_{A_{tk}}(v_{ik})$ to a fuzzy set A_{tk} defined in a rule antecedent for f_k. The inference operation is adopted to compute the firing strength of each rule r_t by using a T-norm (e.g. the min function shown in Eq. (5)) and to derive the overall membership degree of the instance x_i for each label l_j (the presence degree of each type of hate speech) by using one of the T-conorms [11], e.g. the max function shown in Eq. (6).

$$T(\mu_{A_{t1}}(v_{i1}), \mu_{A_{t2}}(v_{i2}), ..., \mu_{A_{td}}(v_{id})) = \min_{k=1}^{d}\{\mu_{A_{tk}}(v_{ik})\} \tag{5}$$

$$S(\mu_{r_1}(x_i), \mu_{r_2}(x_i), ..., \mu_{r_q}(x_i)) = \max_{t=1}^{q}\{\mu_{r_t}(x_i)\} \tag{6}$$

It is proposed to use multiple fuzzy norms (dual pairs of T-norms and T-conorms) for training multiple classifiers. Fuzzification and inference are operated independently using each single classifier. However, the defuzzziciation operation is modified to suit identifying the presence or absence of each hate speech type from a new instance x_i according to a threshold (normally 0.5) of the overall membership degree for each label l_j, i.e. Eq. (7) shows that defuzzification is operated by fusing the m fuzzy classifiers through taking the maximum of the membership degrees $\{\mu_{l_j}^{h_1}(x_i), \mu_{l_j}^{h_2}(x_i), ..., \mu_{l_j}^{h_m}(x_i)\}$ computed through classifiers $\{h_1, h_2, ..., h_m\}$ for each label l_j.

$$\mu_{l_j}^{Ensemble}(x_i) = \max_{f=1}^{m}\{\mu_{l_j}^{h_f}(x_i)\} \tag{7}$$

$$\mu_{l_j}^{Ensemble}(x_i) = \frac{1}{m}\sum_{f=1}^{m}\{\mu_{l_j}^{h_f}(x_i)\} \tag{8}$$

The maximum fusion rule (Eq. (7)) is adopted instead of others such as the mean rule (Eq. (8)), in order to minimize the risk of missing a target type of hate speech to be detected from a tweet, while the type of hate speech is actually present. In other words, if one classifier in a fuzzy ensemble identifies that the membership degree for the target label (hate speech type) is high enough, the fusion of multiple classifiers in the ensemble would make sure the fused membership degree for the target label is high enough. In particular, if the fused membership degree $\mu_{l_j}^{Ensemble}(x_i)$ of instance x_i for a label l_j is greater than the predefined threshold, it would be judged as the case that the target type l_j of hate speech is detected successfully from the instance x_i.

4 EXPERIMENTAL STUDY

In this section, we report an experimental study on identification of hate speech types using four data sets collected for four target types of hate speech, namely: religion, race, disability and sexual orientation. The proposed fuzzy ensemble approach is evaluated by comparing with the state of the art probabilistic approaches such as SVM and DNNs, alongside the use of embedding features. The details on the data sets are described in Section 4.1 and the experimental setup and results are presented in Section 4.2.

4.1 Data

The data sets were collected from Twitter for a period immediately following selected 'trigger' events, which were: for religion, the attack on Lee Rigby in Woolwich, London on 22 May 2013 by Islamist Extremists; for race, the presidential re-election of Barack Obama starting November 6th 2012; for disability, the opening ceremony of the Paralympic games in London, UK on 29th August 2012; and for sexual orientation, the public announcement by Jason Collins on 30th April 2013 - the first active athlete in an American professional sports team to come out as gay. Each event produced datasets between 300,000 and 1.2 million, from which we randomly sampled 2,000 to be human coded. Coders were provided with each tweet and the question: 'is this text offensive or antagonistic in terms of religion/race/sexual orientation/disability?' They were presented with a ternary set of classes - yes, no, undecided.

The results of the annotation exercise produced four 'gold standard' data sets as follows: Religion - 1,901 tweets, with 222 instances of offensive or antagonistic content (11.68% of the annotated sample); Race - 1,876 tweets, with 70 instances of offensive or antagonistic content (3.73% of the annotated sample); Disability - 1,914 tweets, with 51 instances of offensive or antagonistic content (2.66% of the annotated sample); and Sexual Orientation - 1,803 tweets, with 183 instances of offensive or antagonistic content (10.15% of the annotated sample). More details on the data collection and annotation have been presented in [8, 9, 18].

4.2 Experimental Setup and Results

The experiments were conducted by taking the four types of hate speech instances, respectively, from the four single-labelled data sets, to form a training set for evaluating the detection rate of the presence of each hate speech type, through 10-fold cross validation. All the non-hate speech instances were taken from the four data sets to form a holdout test set for evaluating the irrelevance rate on each hate speech type, using the classifiers built on the training set.

For non-DNN based learning methods, all the tweets were pre-processed by converting the words to their lower case, and removing stop words, numbers, punctuation and words that contain less than 3 characters. The pre-processed tweets were then sent for embedding training. Furthermore, embedding features were prepared through adopting distributed bag of words (DBOW). In particular, the learning rate was set to 0.025 with the context window size of 2. Following the embedding learning with the batch size of 10000 for 45 epochs, all the words (appearing at least twice in the corpus) were used for embedding learning to transform each tweet into a document vector with 100 dimensions.

The embedding features were then sent to the learning algorithms for classifiers training. In particular, DT classifiers were trained by using the C4.5 algorithm [27] without pruning. SVM classifiers were trained with using the linear kernel [10]. For GBT training, all the attributes were used and attribute selection for each node of a tree was done by using the same set of attributes. In the setting of the proposed fuzzy ensemble approach, we selected Min/Max norm [37], Product norm [11], Lukasiewicz norm [23] and Yager norm [36], respectively, alongside the border based shrink heuristic as the parameters, for training fuzzy classifiers. The trained fuzzy classifiers were then fused through the maximum rule (Eq. (7)). For defuzzification, a membership degree $\mu_{l_j}(x_i) > 0.5$ is used for identifying the presence of a type of hate speech.

For DNN methods, embedding features were prepared according to the settings of CNN and LSTM reported in [12] and [3], respectively. Based on the prepared embedding features, the classifiers were trained through 2 fully connected layers with 100 units in each layer. The rectified linear unit (ReLU) activation function and the mean squared error (MSE) loss function were used with the learning rate of 0.01, and Stochastic Gradient Descent was used for optimizing the parameters over 20 epochs with the batch size of 10.

The results (rounded to 2 decimal places) on detection rate and irrelevance rate are shown in Tables 1 and 2, respectively. In particular, Table 1 shows that the proposed fuzzy ensemble approach performs the best in terms of the overall detection rate and the rate on the detection of the religion hate speech, whereas it performs the same as or marginally worse than SVM on the detection of the other types of hate speech.

Table 1: Detection rate on four types of hate speech

Learning Method	Overall	Religion	Race	Disability	Sexual Orientation
DT	0.84	0.86	0.70	0.75	0.90
NB	0.63	0.57	0.43	0.73	0.75
SVM	0.92	0.92	**0.84**	**0.86**	0.97
GBT	0.90	0.93	0.80	0.82	0.92
CNN	0.35	0.00	0.00	0.00	**1.00**
LSTM	0.33	0.00	0.00	0.00	0.95
Fuzzy	**0.93**	**0.96**	0.81	**0.86**	0.95

Table 2: Irrelevance rate on four types of hate speech

Learning Method	Overall	Religion	Race	Disability	Sexual Orientation
DT	0.57	0.76	0.58	0.19	0.21
NB	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
SVM	0.37	0.25	0.29	0.20	0.53
GBT	0.54	0.13	0.77	0.12	0.26
CNN	1.00	**0.00**	**0.00**	**0.00**	1.00
LSTM	1.00	**0.00**	**0.00**	**0.00**	1.00
Fuzzy	0.13	0.15	0.00	0.00	0.02

On the other hand, Table 2 shows that the proposed fuzzy ensemble approach outperforms all the other probabilistic approaches, except for NB, in terms of the overall irrelevance rate. In particular, the fuzzy approach performs with an extremely low irrelevance rate on identification of the presence of the race, disability and sexual hate speech (note: 2 non-hate speech instances identified irrelevantly as the race type and 1 identified irrelevantly as the

disability type leading to the rounded irrelevance rate of 0 for both hate speech types) and a fairly low irrelevance rate on identification of religion hate speech (although the rate is considerably higher than the ones on the other types of hate speech).

Although the fuzzy ensemble approach shows a considerably higher overall irrelevance rate than NB, the detection rate obtained using the fuzzy approach is much better than the one obtained using NB. In general, a low detection rate alongside a low irrelevance rate would indicate that the classifier has a strong tendency of not outputting the target class label. In this situation, the irrelevance rate can be extremely low (even equal to 0), since the classifier rarely or even never outputs the target class label.

Table 3: Correlation analysis between different types of hate speech

Label	Religion	Race	Disability	Sexual orientation
Religion	1	-0.221	-0.168	-0.524
Race	-0.221	1	-0.165	-0.221
Disability	-0.168	-0.165	1	-0.237
Sexual Orientation	-0.524	-0.221	-0.237	1

On the basis of the above discussions, the overall performance of the proposed fuzzy ensemble approach is better than the ones of the other probabilistic approaches. In addition, the fuzzy approach can also be used for effectively identifying the label correlations, where the probabilistic approaches fall short [30, 39]. The results on correlation analysis between different labels are shown in Table 3, which indicates the correlation coefficient for each pair of class labels in terms of the membership degrees.

In general, there are three extreme cases, namely, 'mutually exclusive' (the most strongly negative correlation), 'independent' and 'identical' (the most strongly positive correlation), when the correlation coefficients are -1, 0 and +1, respectively. The results shown in Table 3 indicate that the four types of hate speech generally involve very weakly negative correlations. The results on correlation analysis between different types of hate speech show supporting evidence that it is much more necessary to formulate the problem as multi-label classification instead of single-label one. In order to show the correlation analysis in more depth, we conduct an extended study on topic detection using the Latent Dirichlet Allocation (LDA) algorithm. The results are shown in Table 4.

In general, the results show high diversity of hate speech instances, i.e. diverse topics involved in hate speech instances. Some common topics can be found from different hate speech data sets. For example, both religion and race hate speech data involve the topics on black and white people. Also, the topics on christian can be found from both religion and sexual orientation hate speech data. In addition, sports related topics are involved in both disability and sexual orientation hate speech data.

The results on topic detection generally indicate that hate speech instances are highly diverse, leading to a high likelihood that a single tweet could be related to more than one type of hate speech, although the class labels may not have strongly positive correlations. In this case, the proposed fuzzy approach is even more suitable for dealing with the diversity, since it is essentially an instance based approach of fuzzy rule learning. In other words, the generation of each fuzzy rule is triggered by an instance without direct discrimination between different classes.

Table 4: Topics detected from hate speech instances in four data sets using LDA

Dataset	Topic ID	Word Cluster
Religion	0	nigger, paki, black, guy, incident
	1	black, people, shot, killed, white
	2	black, edl, nigga, home, send
	3	niggers, nigga, shame, kill, fuck
	4	black, niggas, paki, nigger, islam
Race	0	nigga, niggas, america, shit, won
	1	black, voted, people, tweets, cuz
	2	white, black, romney, won, gotta
	3	niggas, yall, president, hate, nigga
	4	nigga, real, mitt, house, white
Disability	0	irony, leg, called, lickmynippless, mybad
	1	dont, swimmer, ill, understand, drunk
	2	han, labradors, cancelled, highjumps, blind
	3	moving, arsed, wish, falling, women
	4	wheelchair, team, day, events, frankieboyle
Sexual Orientation	0	gay, chris, broussard, http, dick
	1	gay, hes, nba, call, ball
	2	gay, aint, niggas, pretty, espn
	3	nba, faggot, hes, gay, coming
	4	media, tebow, tim, christian, yourself

5 CONCLUSIONS

In this paper, we have proposed a theoretical framework that involves fuzzy mutli-task learning for hate speech types detection. In particular, a novel formulation of the problem has been introduced in the setting of semi-supervised multi-label learning from single-labelled data. An experimental study has been reported to evaluate the performance of the proposed fuzzy ensemble approach using the two new metrics referred to as 'detection rate' and 'irrelevance rate', respectively. The experimental results show that the proposed fuzzy approach outperforms the state of the art probabilistic approaches such as SVM and DNNs on embedding features. Also, the proposed fuzzy approach provides an intensity score for the presence of each type of hate speech from a tweet and enables the analysis of correlation between different labels, while the probabilistic approaches fall short in this aspect.

In the future, we will collect more diverse types of hate speech instances to increase the number of labels, and investigate the impacts of the increased number of labels on the detection performance of each type of hate speech as well as the intersectionality between different types. Also, an extension of the currently proposed fuzzy approach to involve multiple iterations of semi-supervised multi-labelling would be another interesting further direction in the setting of self-labelling based semi-supervised learning [20], i.e. the original single-labelled data becomes multi-labelled for further training after the first iteration of training.

ACKNOWLEDGMENTS

This paper is supported partially by Economic and Social Research Council (Grant Number: ES/P010695/1) and partially by the RAND Corporation (Award Number: 2016-MU-MU-0009).

REFERENCES

[1] Zahra Ahmadi and Stefan Kramer. 2018. A label compression method for online multi-label classification. *Pattern Recognition Letters* 111 (2018), 64–71.

[2] Wafa Alorainy, Pete Burnap, Han Liu, Amir Javed, and Matthew L. Williams. 2018. Suspended Accounts: A Source of Tweets with Disgust and Anger Emotions for Augmenting Hate Speech Data Sample. In *International Conference on Machine Learning and Cynernetics*. IEEE, Chengdu, China.

[3] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. ACM, Perth, Australia, 759–760.

[4] Chara Bakalis. 2018. Rethinking cyberhate laws. *Information & Communications Technology Law*, 27, 1 (2018), 86–110.

[5] James Banks. 2010. Regulating hate speech online. *International Review of Law, Computers and Technology* 24, 3 (2010), 233–239.

[6] James Banks. 2011. European regulation of cross-border hate speech in cyberspace: The limits of legislation. *European Journal of Crime, Criminal Law and Criminal Justice* 19, 1 (2011), 1–13.

[7] Michael R. Berthold. 2003. Mixed Fuzzy Rule Formation. *International Journal of Approximate Reasoning* 32 (2003), 67–84.

[8] Pete Burnap and Matthew Williams. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy and Internet* 7, 2 (2015), 223–242.

[9] Pete Burnap and Matthew Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5, 11 (2016).

[10] Nello Cristianini. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge.

[11] Thomas R. Gabriel and Michael R. Berthold. 2004. Influence of fuzzy norms and other heuristics on Mixed fuzzy rule formation. *International Journal of Approximate Reasoning* 35 (2004), 195–202.

[12] Bjorn Gamback and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-Speech. In *1st Workshop on Abusive Language Online*. ACM, Vancouver, Canada.

[13] Tieliang Gong, Guangtao Wang, Jieping Ye, and Zongben Xu. 2018. Margin Based PU Learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. AAAI, New Orleans, Louisiana, USA, 3037–3044.

[14] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press, Bellevue, Washington, 1621–1622.

[15] Xiaoli Li and Bing Liu. 2005. Learning from Positive and Unlabeled Examples with Different Data Distributions. In *European Conference on Machine Learning*. Springer, Porto, Portugal, 218–229.

[16] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip Yu. 2003. Building Text Classifiers Using Positive and Unlabeled Examples. In *Proceedings of the Third IEEE International Conference on Data Mining*. IEEE, Melbourne, Florida, USA, 1–8.

[17] Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. 2002. Partially Supervised Classification of Text Documents. In *Proceedings of the Nineteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers, Sydney, Australia, 387–394.

[18] Han Liu, Pete Burnap, Wafa Alorainy, and Matthew L. Williams. 2019. A Fuzzy Approach to Text Classification with Two Stage Training for Ambiguous Instances. *IEEE Transactions on Computational Social Systems* 6 (2019).

[19] Han Liu and Mihaela Cocea. 2017. Fuzzy Rule Based Systems for Interpretable Sentiment Analysis. In *International Conference on Advanced Computational Intelligence*. IEEE, Doha, Qatar, 129–136.

[20] Han Liu and Mihaela Cocea. 2018. *Granular Computing Based Machine Learning: A Big Data Processing Approach*. Springer, Berlin.

[21] Han Liu, Mihaela Cocea, and Weili Ding. 2018. Multi-Task Learning for Intelligent Data Processing in Granular Computing Context. *Granular Computing* 3, 3 (2018), 257–273.

[22] Han Liu, Mihaela Cocea, Alaa Mohasseb, and Mohamed Bader. 2017. Transformation of Discriminative Single-Task Classification into Generative Multi-Task Classification in Machine Learning Context. In *International Conference on Advanced Computational Intelligence*. IEEE, Doha, Qatar, 66–73.

[23] Jan Lukasiewicz. 1970. *Selected WorksâĂŞâĂŞStudies in Logic and the Foundations of Mathematics*. North-Holland Publishing, Amsterdam.

[24] Irene Nemes. 2010. Regulating Hate Speech in Cyberspace: Issues of Desirability and Efficacy. *Journal of Information and Communications Technology Law* 11, 3 (2010), 193–220.

[25] Chikashi Nobata, Joel Tetreault, and Achint Thomas. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*. ACM, Montreal, Quebec, Canada, 145–153.

[26] Ji Ho Park and Pascale Fung. 2017. One-step and Two-step Classification for Abusive Language Detection on Twitter. In *1st Workshop on Abusive Language Online*. ACM, Vancouver, Canada.

[27] Ross J. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco.

[28] Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive Language Detection Using Multi-level Classification. In *Proceedings of the 23rd Canadian conference on Advances in Artificial Intelligence*. Springer, Ottawa, Canada, 16–27.

[29] Jesse Read. 2013. Multi-label Classification. In *The Second School on Machine Learning and Knowledge Discovery in Databases*. Brazilian Computer Society, Sao Carlos, Brazil.

[30] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning* 85 (2011), 333–359.

[31] KÃűffer Sebastian, Riehle Dennis M, HÃűhenberger Steffen, and Becker JÃűrg. 2018. Discussing the Value of Automatic Hate Speech Detection in Online Debates. In *Multikonferenz Wirtschaftsinformatik*. GITO-Verlag, Leuphana, Germany, 83–94.

[32] G. Tsoumakas and I. Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* 3, 3 (2007), 1–13.

[33] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of NAACL-HLT 2016*. ACL, San Diego, California, USA, 88–93.

[34] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access* PP, 99 (2018), 1–11.

[35] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. Springer, Maui, Hawaii, USA, 1980–1984.

[36] Ronald R. Yager, S. Ovchinnikov, R. M. Tong, and H. T. Ngugen. 1987. *Fuzzy Sets and Applications*. Wiley, New York.

[37] Lotfi Zadeh. 2015. Fuzzy Logic: A Personal Perspective. *Fuzzy Sets and Systems* 281 (December 2015), 4–20.

[38] Bangzuo Zhang and Wanli Zuo. 2008. Learning from Positive and Unlabeled Examples: A Survey. In *2008 International Symposiums on Information Processing*. IEEE, Moscow, Russia, 650–654.

[39] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (2014), 1819–1837.

[40] Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *European Semantic Web Conference*. Springer, Heraklion, Crete, 745–760.