HATE SPEECH DETECTION OF ARABIC SHORTTEXT

Abdullah Aref[1], Rana Husni Al Mahmoud[2],
Khaled Taha[3] and Mahmoud Al-Sharif[3]

[1]Computer Science Department, Princess Sumaya
University for Technology, Amman, Jordan
[2]Computer Science Department, University of Jordan, Amman, Jordan
[3]Social Media Lab, Trafalgar AI, Amman, Jordan

ABSTRACT

The aim of sentiment analysis is to automatically extract the opinions from a certain text and decide its sentiment. In this paper, we introduce the first publicly-available Twitter dataset on Sunnah and Shia (SSTD), as part of a religious hate speech which is a sub problem of the general hate speech. We, further, provide a detailed review of the data collection process and our annotation guidelines such that a reliable dataset annotation is guaranteed. We employed many stand-alone classification algorithms on the Twitter hate speech dataset, including Random Forest, Complement NB, DecisionTree, and SVM and two deep learning methods CNN and RNN. We further study the influence of word embedding dimensions FastText and word2vec. In all our experiments, all classification algorithms are trained using a random split of data (66% for training and 34% for testing). The two datasets were stratified sampling of the original dataset. The CNN-FastText achieves the highest F-Measure (52.0%) followed by the CNN-Word2vec (49.0%), showing that neural models with FastText word embedding outperform classical feature-based models.

KEYWORDS

HateSpeech, Dataset, Text classification, Sentiment analysis.

1. INTRODUCTION

Hate speech is a crime that has been growing in recent years, not only in face-to-face interactions but also in online communication [1]. Social media platforms allow users to broadcast any sort of messages in these systems and to reach millions of users in a short period and at near zero cost [2]. The freedom available to social media users to express their opinions and the anonymity provided by these environments [1] made it easier to spread hate propaganda against individuals or groups [3], [4]. This provoked the need for automatic detection of hate speech contents shared across social media platforms [3], especially if such online contents can direct physical hate crimes [3].

The problem in hate speech is wide in nature and varies according to the type of hate speech (sexism, racism, religious hate speech, etc..).The absence of human annotated vocabulary that explicitly reveals the presence of hate speech, makes the available hate speech corpora sparse and noisy [5]. Even though, many studies have been conducted on automatic detection of hate speech,

David C. Wyld et al. (Eds): ITCSE, NLCA, ICAIT, CAIML, ICDIPV, CRYPIS, WiMo - 2020
pp. 81-94, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.100507

only a few of them can result in high precision and recall rates [6], and the tools provided are scarce[1].

When compared to English, Arabic is considered an under-resourced language. The complexity and richness of Arabic morphology combined with the existence of different dialects add up more challenges to Arabic NLP research [7]. Despite the existence of many researches that investigated anti-social behaviours such as, abusive or offensive language and cyberbullying, a limited number of researches have contributed to hate speech detection in Arabic [6].

In this work, we address the hate speech between Sunnah and Shia, as part of the religious hate speech which is a sub problem of the general hate speech. In the absence of a labelled data-set for this purpose, we create Sunnah Shia Twitter Dataset (SSTD) analyse it using various well known machine learning approaches.

2. RELATED WORK

Several studies have been conducted with the goal of describing online hate speech and which groups are more threatening. Descriptive statistics about hate speech can be found in the literature [1]. including Racism [8], Sexism [9], Prejudice toward refugees [10], Homophobia [11], and General hate speech [2]. Other researchers focused on algorithms for hate speech detection and used text mining and machine learning for hate speech classification [1] such as [3] and [7]. Many text mining strategies have been adapted for automatic detection of hate speech[1].

Features representation for hate speech detection including distance metric are addressed in [12], dictionaries, Bag-of-words (BOW) [8], [13], N-grams [14], Profanity Windows [15], TF-IDF [16], part of-speech [17], Lexical Syntactic Feature-based [18], Rule Based [19], Topic Classification [20], Sentiment [21], Word Embeddings [22], Typed Dependencies [17], and Deep Learning [23]. A more in depth survey for features representation for hate speech detection can be found in [1]. In the litterateur, supervised, semi-supervised and unsupervised approaches machine learning classification algorithms were used for hate speech detection [6].

Deep learning models showed promising future text sentiment analysis [24]. Recurrent Neural Networks (RNN) were used in [25] with word frequency vectorization to implement the features. Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), and FastText, combined with numerous features like TFIDF and Bag of Words (BoW) vectors were used in [26] to detect racism and sexism, including. LSTM with random embeddings found to outperform other approaches [27].

Levantine Hate Speech and Abusive (L-HSAB) Twitter dataset introduced in [3] and two well known machine learning algorithms were evaluated. The results indicated the the multinomial NB outperformes SVM [3].

The problem of religious hatred in Arabic twitter was tackled in [7], and various deep learning algorithms were tested for this task including GRU RNN which were found to work better than LSTM with smaller datasets [6].

This work has two contributions: we address hate speech between Sunnah and Shia, in twitter and we created SSTD and applied various machine learning approaches on it.

3. WHAT IS HATE SPEECH?

A universal definition of Hate Speech was found to be difficult to derive. There are a variety of definitions by a number of international and organizations like the United Nations and the European Union, in addition to key NGO activist organizations. A survey of multiple definitions from different origins were presented in [1]. By studying common factors across most popular definitions and implementing practical, philosophical, cultural and technological considerations, a definition of Hate Speech was derived for the purpose of this study. It can be stipulated as follows:

" Any Pejoration or Threat Directed at a Group of Protected Attributes".

Pejoration is articulated to be one or any combination of: a) expression or incitement to disdain b) expression or incitement to insult c) expression of belittling false generalization.

Threat is articulated to be one or any combination of: a) expression or incitement to violent action b) expression or incitement to isolate c) expression or incitement to hate.

4. METHODOLOGY

In this work we follow a methodology of six stages namely. 1) keywords selection, 2) collecting the dataset, 3) prepossessing 4) feature extraction, 5) model development, and 6) evaluation. Figure 1 shows the flow layout of these stages. The following subsections discuss each stage in more details.

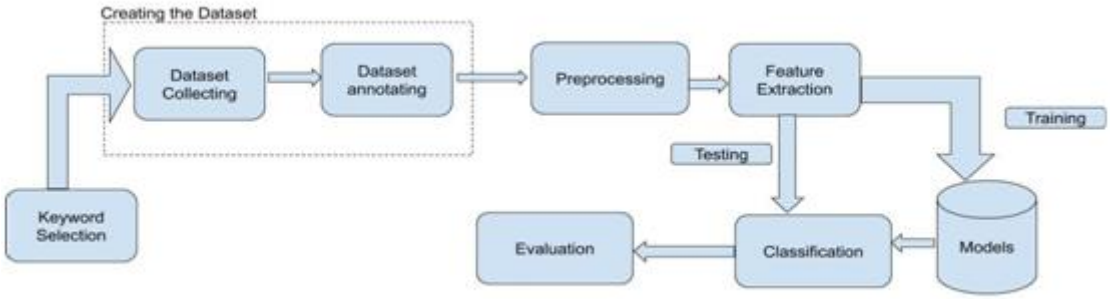

Figure 1. Methodology

4.1. Keywords selection

In doing research on the use of Twitter for hate speech targeting Sunnah and Shia, the first challenge is to collect a comprehensive (or at least representative) sample of tweets for the topic. To tackle this issue, one simple and straightforward solution is to concentrate on tweets that contain words that are likely to indicate hate speech combined with words that are likely to indicate one of the two groups

We created a set of words to work as our seeding list. The list words extracted from a set of the hate speech related documents [28]–[31]. There are around 365 words, appearing in these documents that were manually assessed and scored by one member of the team for being hate or not and the degree of hatred in these keywords. The selected keywords that had high score are presented in Table I. Keywords that related to Sunnah and Shia topic presented in Table II.

Table I: Keywords related to Sunnah and Shia with high score

Keyword	**Transliteration**	**Keyword**	Transliteration	**Keyword**	**Transliteration**
اجتثاث	Egtithath	دحابشه	Dahabshah	أخونجي	Ekhwanji
إرهابي	Erhabe	دواعش	Dawaesh	العدوان	Al Edwan
الحقد الشيعي	Al-Heqd Al Sha'abe	رخيص	Rakhees	العن اليهود	Aleyn AlYahood
الحوثيون	Al Hotheon	زنادقة	Zanadeka	اللهم العن	AllahomAleyn
السنة النواصب	Al Sunnah Al Nawaswb	سعود	Souod	المشركين	Al Mushreken
الشرك	Al Sherk	شراميط	Sharameed	ايران	Iran
الشيعة ليسوا مسلمين	Al Sheialiso Muslimeen	شيعه	Sheia	الشيعة حمير اليهود	Al Sheia Hamer Al Yahood
العر بان	Al Orban	ضد	Ded	أسيادك	Asyadak
القردة والخناز ير	Al kerada and al khanazeer	عبيد	Abeed	أهل البدع	Ahl Albeda'a
المد الإ يراني	Al Mad Al Irani	عملاء	Omala'a	أوغاد	Awgad
انفصالي	Enfesali	فاسدون	Fasedoon	آل سلول	Al Salol
أخذ عز يز مقتدر	Akhdaazizmuqtader	فاسقون	Faseqon	حاقد	Haked
أذرع	Athroey	قتل	Katal	حقارة	Haqara
أعداء	Aeyda'a	كافر	Kafer	حقير	Haqeer
أو باش	Awbash	كلاب	Kelab	خايس	Khaes
أولاد حرام	Awlad Al haram	لعن	Laa'an	خرا	Khara
تبا لكم	Taba'a Lak	لعنه	Laanah	خرى	Khara
حثالة	Huthala	ليبرالي	Lebraly	خونة	Khawana
حقد	Heked	مخنث	Mukhanath	شيطان	Sheitan
حمير	Hamer	ملحد	Mulhed	صفوي	Safawy
خائن	Khaen	نجس	Najas	عاهرة	Aahera
خرة	Kharah	نواصب	Nawaseb	عصابة	Aesaba
خوارج	Khawarej	وهابي	Wahabe	عميل	Ameel
خيانة	Kheana	يهودي	Yahodi	فاسدين	Fasedeen
إرهاب	Erhab	دحباشي	Dehbashi	فساد	Fasad
الإرهابيون	Al Erhabeon	رافضه	Rafeda	قذر	Kather
الحوثي	Al Hothe	زق	Zaq	كفر	Kafar
السنة الدواعش	Al Suna Al Dawaesh	سراق	Suraq	كلب	Kalb
السنة إرهابية	Al Sunaerhabia	سفلة	Safala	لعنة اللهّٰ	La'anat Allah
منافقون	Munafekon	وهابية	Wahabiah	لعين	Laeen
نصراني	Nasrane	و بئس المصير	Wabeas Al maseer	مجوس	Majoos
مرتزقة	Murtasaqah				

4.2. Creating the Dataset

The process of creating the dataset consists of two phases, collecting and annotating the dataset.

1) Dataset collection: Twitter offers several open, public data access options. Every approach has particular advantages and limitations. The researcher's aims to determine which approach is effective in a given context.

The standard Twitter APIs consist of REST APIs and Streaming APIs[1]. Twitter provides Representational State Transfer (REST) search API for searching tweets from Twitter's search index. REST API provides seven days historical results. While the streaming API gives results from the point of the query. Streaming API can be used to track a specific query in real-time. Twitter search API has many limitations as mentioned on their web site[2]. We created a number of queries that contain all concatenations of the hate speech keywords and group keywords of Tables I & II. We carried two searches, one on 30/9/2019 and the other on 5/10/2019. Due to twitter limitations in search API, it retrieved around 8220 tweets within the 7 days before search date. Our dataset tweets are date between 23/9/2019 and 5/10/2019.

To make our models, we select a dataset that contains 3235 tweets. The selected dataset was stratified sampling of the original dataset. Which means that each combination of the topic and groups tweets has the same ratio in sampled and original dataset.

Table II: Group keywords

Group	Transliteration	Group	Transliteration
الشيعة	Al Sheia	رافضي	Rafede
أهل السنة	Ahl Al Sunah	السلفيون	Al Salafeyon
سني	Suni	روافض	Rawafed
شيعي	Sheie	قتلة الحسين	Katalat Al Husien
وهابية	Wahabia	مجوس	Majos
سلفي	Salafi	مذهبي	Mathhabe
شيعة	Sheiah	صفوي	Safawe
طائفي	Taeefe	أبناء المتعة	Abna'a Al Muteyah
فرس	Furs	وهابي	Wahabe

2) Evaluation rules: The following evaluation rules were derived for improving the quality of labelling the individual tweets.

- Criticism directed at political regimes or states, is not to be considered hate speech, even if it was severe.
- Special consideration is applied to the context of insults directed at individuals, since some of it can be interpreted as hate speech.
- Consideration of the variety of grounds that Hate Speech is based upon across extended geographies (example: Arab West: Islamic vs. Secular, Arab East: Sunni vs. Shai).
- New terms that appear to be neutral in normal contexts, are found to be extremely pejorative against specific groups, are added to stop words list of hate speech.

- Cursing, is still a common practice in the Arab culture (as opposed to Western culture). It's important to distinguish between cursing as a common practice from cursing as a hate speech.

- Irony, metaphors and figurative speech can be used as a maneuver around hate speech, especially in countries that have strict legal liabilities against it.

- Special attention is paid to the religious terms that carry meanings of supplication to God, exclamation or expression of weakness - since all can be used as a religious cover to hate speech.

- Pejoration directed at women, even if done on cultural or social basis, is labeled as so with no leniency

- Special attention is paid in order not to constraint freedom of speech in the attempt to alleviate hate speech.

- Stigmatizing terms are found to be extremely pejorative to groups, and can't be interpreted in any positive form, hence they are labeled as hate speech wherever it appears (example: "Rawafid" for Shiis, "Irhabi" for Sunnis).

- Wherever a new word is found to be associated with Western Arab's hate speech, it is added to hate speech stop words.

- Political correctness is not to be mistaken with hate speech.

- Sole usage of controversial historical topics, is not to be considered hate speech. It must be associated with terms or incitments that indicate clear hate speech.

- Citation of hate speech, is not to be considered or labeled as hate speech.

- Extreme criticism to protected groups is to be considered non hate, if it's done only in generalized political context.

- In Arab culture, it's extremely important not to take everything at "face value", and dig down into understanding roots and deeper meanings of expressions used.

3) Data Annotation: We assigned the labelling task to two annotators; Khaled Taha[3], and MamoudAl Sharief[4].

We asked the annotators to judge each tweet and categorize them as either contains hate speech (HATE); or does not contain hate (Not Hate). The agreement between the two annotators was 0.85%. In a corpora for Hate Speech, annotator disagreement can be related to the fact that there are many rules to be applied on the tweet to indicate it as hate tweet. Therefore, any tweet is considered to be Hate if it is marked as Hate by at least one of the annotators. It is not uncommon to discard tweets with high disagreements among annotators. While some claim that this is done to remove noise from low annotator quality; this argument does not hold when considering that high rates of consensus annotator agreement are present in these datasets. This indicates that the issue is not weak annotators, but rather difficult data that are not in the predefined categories [32].

The resulting dataset contains the text of each tweet along with the adjudicated label. The distribution of the texts across the two classes is shown in Table III.

Table III: Hate Speech Identification dataset classes

Class Label	Number of tweets
Not Hate	2590
Hate	642

4.3. Data pre-processing

The text pre-processing phase includes a set of processes applied on the dataset. They mainly include: normalization of some Arabic letterforms, tokenization of words, stop-words removal and stemming.

1) **Normalization**: Generally, text pre-processing tasks attempt to reduce the noise using normalization. In this work, we employed the following normalization steps:

   - Remove non letters and special characters ($,&,%,...)
   - Remove non Arabic letters
   - Replace initial آ , إ or أ with bare alef ا
   - Replace final ة with ه
   - Remove ال from the beginning of a word
   - Replace final ي with ى

2) **Tokenization**: This step is used to analyze text linguistically. It breaks strings of characters, words, and punctuation marks into tokens during the indexing process.

3) **Stop Words removal**: Words that do not affect the meaning of the text usually referred to as Stop Words, such as prepositions. Every natural language has its own list of stop-words.

4) **Morphological analysis and stemming:** Arabic morphological analysis and root extraction are essentials for many Arabic applications such as information retrieval and data mining. In the literature adequate works tackling the problem of Arabic morphological analysis is given in [33]–[35]. Because of its nature, Arabic found to be very difficult to stem [36]. Mainly, there are two kinds of stemming algorithms in Arabic: a root-based approach, for example Khoja and Garside [37]; and stem-based (light stemming) approach [38]. In this work we apply light stemmer on the text.

4.4. Feature extraction

Feature extraction is a pre-processing step toward knowledge discovery and dimensionality reduction. In this stage, features (i.e. POSTs) were extracted from documents based on their calculated weights in the collection. In the literature, several features- extraction approaches were used [39]. In this work, we used the TF-IDF as a weighting scheme for feature selection. TF-IDF is used to determine the keywords that can identify or categorize some specific documents in a

collection. TF-IDF attempts to combine both the number of times the word t occurs in document d, referred to as TF(t,d), with the inverse document frequency, referred to as IDF(t) [40].

For each of the deep neural networks methods, we initialize the word embeddings with either Word2Vec embeddings or FastText embeddings that give better results compared with GloVe embedding.

For the Word2Vec and FastText model, it has been trained on SSTD using the skipgram model where the context window size was set to (10) , and the vector size wasset to (100). In addition to wordNgrams equals 6 for FastText model.

4.5. Evaluation measures

We compare the performance of all these methods based on a set of standard evaluation measurements (described next) with respect on the confusion matrix shown in Table 3. We have two class labels in the dataset namely Hate, Not Hate. The four possible outcomes of the confusion matrix are as follows:

A : A Not Hate Tweet is correctly classified as a Not Hate

B : A Not Hate Tweet is incorrectly classified as Hate

C : A Hate Tweet is incorrectly classified as a Not Hate

D : A Hate Tweet is correctly classified as a Hate

Table 3: Confusion Matrix

		Predicted	
		Not Hate	Hate
Actual	Not Hate	A(TN)	B(FP)
	Hate	C(FN)	E(TP)

Following is a description of the evaluation measures used to compare the performance of the different classification methods used in this study:

Basic measures

– Accuracy: Accuracy is a metric used to estimate how a classifier can correctly predict Hate, Not Hate instances for each class. It can be calculated as the ratio of correctly classified instances to the total number of instances, as given in Eq. 1 which is adapted from the general accuracy equation (32).

$$\text{Acuuracy}=\frac{A+D}{A+B+C+D} \quad (1)$$

– Precision: Precision of a class C, where C is Not Hate, or Hate is the ratio of the correctly predicted to the total predicted samples and is calculated as in Eq. 2 which is adapted from the general macro precision equation (32).

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (2)$$

– Recall: Recall of a class C, where C is Not Hate, or Hate is the ratio of C instances that are correctly predicted to the total number of actual C instances. It can be calculated as in Eq. 3 which is adapted from the general macro recall equation (32).

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (3)$$

– F-Measure: F-Measure is a composition of Precision and Recall. It is a consistent average of the two metrics which is used as an accumulated performance score. F-Measure of a class C, where C is Not Hate, or Hate can be calculated as in Eq. 4 which is adapted from the general macro F-Measure equation (32).

$$\text{F-Measure} = \frac{2*(precision*Recall)}{(Precision+Recall)} \quad (4)$$

Mainly we used two metrics to evaluate the performance of the developed classification model, namely, the precision and recall which can be summarized in F-measure, which is commonly used in the literature for imbalanced datasets as the accuracy measure is not of interest is similar cases.

5. EXPERIMENTS AND EVALUATION RESULTS

This section presents the experimental analysis of the performance of several well known classifiers over the created dataset.

5.1. Experiments setup

All experiments were conducted using a personal computer with Intel® core i5-5500U CPU @ 2.53GHz / 8 GB RAM. To conduct the experiments, we used Python and Anaconda framework. The Scikit-learn library was selected to implement the classification and to measure the machine learning algorithms' performance; we applied neural network classifiers using the Keras Python library.

5.2. Experiments

We evaluated the performance of several machine learning and deep learning algorithms including Random Forest [41], Complement NB [42], Decision Tree [43], support Vector Machine (SVM) [44], Convolutional Neural Network (CNN) [45], and Recurrent Neural Networks (RNN) [46].

In all experiments, all classification algorithms are trained using a random split of data (66% for training and 34% for testing). The two datasets were stratified sampling of the original dataset.

The testing dataset is unseen during training the model and the performance of the model is determined by predictions applied on the testing dataset.

Text of tweets used for our analysis of the first four classifiers, and TF-IDF used for feature extraction generate numerical features with bag-of-words strategy. We further study the influence of word embedding dimensions FastText [47] and word2vec [48] on deep learning algorithms; Manley CNN and RNN,

Table V: Classification results over Sunnah Shia Dataset. Best values are in bold typeface

	accuracy	**Precision**		**Recall**		**F-Score**	
		0	1	0	1	0	1
Random Forest	0.78	0.78	**0.78**	**0.99**	0.086	0.87	0.15
Complment NB	0.68	**0.85**	0.39	0.72	0.579	0.78	0.46
DecisionTree	0.74	0.84	0.46	0.83	0.46	0.83	0.46
SVM	0.8	0.81	0.64	0.95	0.29	**0.88**	0.4
cnn	0.8	0.83	0.62	0.93	0.39	**0.88**	0.48
Rnn	0.8	0.83	0.63	0.93	0.37	**0.88**	0.47
CNN+fasttext	0.71	0.88	0.42	0.71	**0.69**	0.79	**0.52**
Rnn+ fasttext	0.76	0.83	0.49	0.87	0.4	0.85	0.44
Cnn+word2vec	**0.79**	0.84	0.59	0.91	0.42	0.87	0.49
Rnn +word2vec	0.74	0.83	0.44	0.84	0.42	0.83	0.43

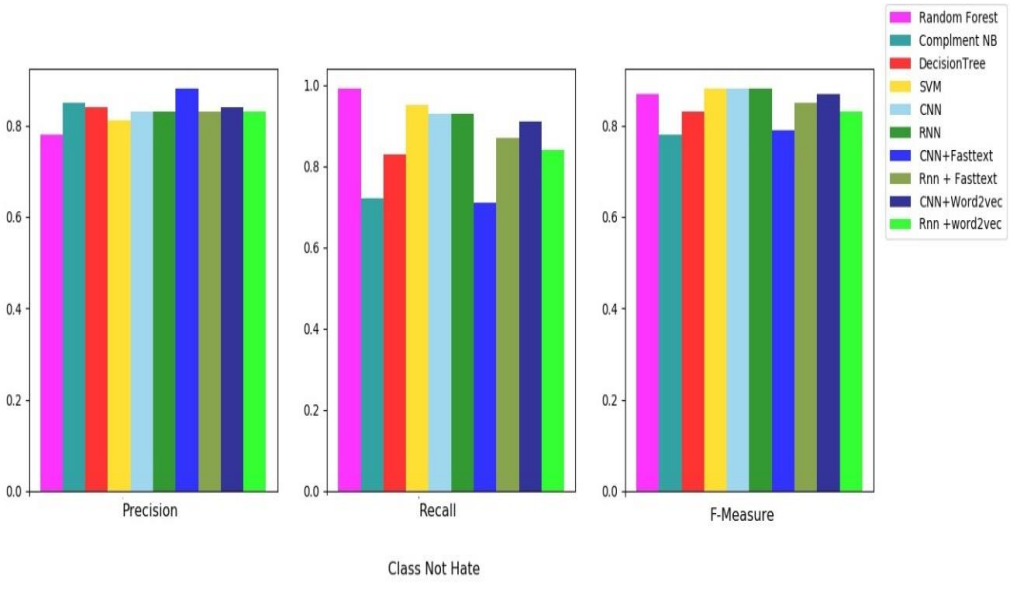

Figure3. The calculated measures for class Not Hate of all tested models

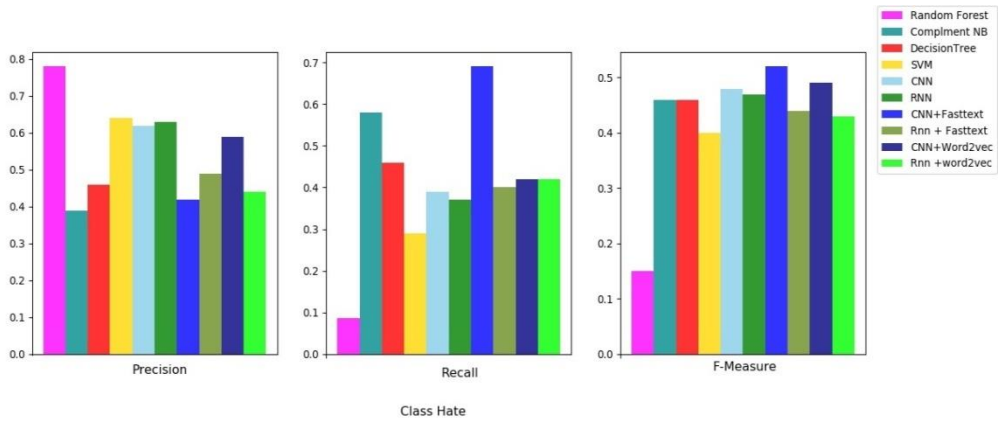

Figure4. The calculated measures for class Hate of all tested models

Experimental results are shown in Table V. As we are targeting the minority class improvement in learning from imbalanced data distributions, it is more important to improve the F-Measure for each class, than improving accuracy. Thus, we studied the behaviour of F-Measure in experiments. We can see from Table V that deep learning models based on CNN and RNN outperformed other models. Also we find that the use of FastText combined with CNN and RNN, outperforms CNN and RNN alone or combined with Word2vec.

Figure 3 and Figure 4 shows the overall performance of ten classification models tested over SSTD. They are reported in terms of Precision, Recall and F1-measure

Our first observation from Figure 3, is that there is no majority difference in precision for class "NOT Hate". Random Forest classifier outperformed all other models in recall measure. The highest F-Measure values achieved for "Not Hate" class when applying SVM, RNN and CNN models.

As presented in the Figure 4, CNN outperformed other algorithms with respect to F-Measure for Class "Hate". This agrees with the findings in [49] that CNN is a powerful tool to improve the prediction performance. CNN's first success in sentiment analysis was triggered by research on document classification [50], where CNN has demonstrated state-of-the-art results in document classification datasets, this performance has led to a rise in deep neural network sentiment analysis research [50].

The benefit of the FastText feature over Word2Vec is that it integrates subword information into the embedding learning process. Through the combination of learned ngram embeddings it can learn similar embedding for words sharing a common stem and also generate embedding for unseen words into the test set [51].

6. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we introduced SSTD, a dataset targeting religious hate speech (Sunnah and Shia). To build the dataset, we retrieved many tweets from Twitter, and asked 2 annotators to manually label the tweets following a set of agreed on rules. The dataset combined 3232 tweets with 2 categories: "Hate" and "Not Hate". As hate speech annotation rely on several rules as well as the annotators' knowledge, experience, and assumptions, the agreement between annotators remains an issue. The performance of several well known machine learning and deep learning algorithms were analysed using the SSTD. The results indicated the outperformance of CNN over other

tested algorithms. A natural future step would involve building publicly-available datasets for hates speech targeting other Muslim groups such as Sufi and Muslim brotherhood as well as that targeting other religious groups such as The Copts and Orthodox and so on.

REFERENCES

[1] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Computing Surveys (CSUR), vol. 51, no. 4, p. 85, 2018.

[2] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the targets of hate in online social media," in Tenth International AAAI Conference on Web and Social Media, 2016.

[3] H. Mulki, H. Haddad, C. B. Ali, and H. Alshabani, "L-hsab: A levantine twitter dataset for hate speech and abusive language," in Proceedings of the Third Workshop on Abusive Language Online, 2019, pp. 111–118.

[4] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," Aggression and violent behavior, vol. 40, pp. 108–118, 2018.

[5] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," arXiv preprint arXiv: . , 2019.

[6] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: a survey on multilingual corpus," Computer Science & Information Technology (CS & IT), vol. 9, no. 2, p. 83, 2019.

[7] N. Albadi, M. Kurdi, and S. Mishra, "Are they our brothers? analysis and detection of religious hate speech in the arabictwittersphere," in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2018, pp. 69–76.

[8] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in Twenty-seventh AAAI conference on artificial intelligence, 2013.

[9] A. Jha and R. Mamidi, "When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data," in Proceedings of the second workshop on NLP and computational social science, 2017, pp. 7–16.

[10] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the reliability of hate speech annotations: The case of the european refugee crisis," arXiv preprint arXiv: . , 2017.

[11] V. Reddy, "Perverts and sodomites: Homophobia as hate speech in africa," Southern African Linguistics and Applied Language Studies, vol. 20, no. 3, pp. 163–175, 2002.

[12] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proceedings of the second workshop on language in social media. Association for Computational Linguistics, 2012, pp. 19–26.

[13] P. Burnap and M. L. Williams, "Us and them: identifying cyber hate on twitter across multiple protected characteristics," EPJ Data Science, vol. 5, no. 1, p. 11, 2016.

[14] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[15] M. Dadvar, F. d. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR). University of Ghent, 2012.

[16] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in fifth international AAAI conference on weblogs and social media, 2011.

[17] P. Burnap and M. L. Williams, "Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making," 2014.

[18] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in International Conference on Privacy, Security, Risk and Trust and International Confernece on Social Computing.
IEEE, 2012, pp. 71–80.

[19] Y. Haralambous and P. Lenca, "Text classification using association rules, dependency pruning and hyperonymization," arXiv preprint arXiv: . , 2014.

[20] S. Agarwal and A. Sureka, "Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website," arXiv preprint arXiv: . , 2017.

[21] S. Liu and T. Forss, “New classification models for detecting hate and violence web content,” in th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC K), vol. 1. IEEE, 2015, pp. 487–495.
[22] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings,” in Proceedings of the th international conference on world wide web. ACM, 2015, pp. 29–30.
[23] S. Yuan, X. Wu, and Y. Xiang, “A two phase deep learning model for identifying discrimination from tweets.” in EDBT, 2016, pp. 696–697.
[24] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1253, 2018.
[25] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, “Effective hate-speech detection in twitter data using recurrent neural networks,” Applied Intelligence, vol. 48, no. 12, pp. 4730– 4742, 2018.
[26] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in Proceedings of the th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017, pp. 759–760.
[27] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, “A unified deep learning architecture for abuse detection,” in Proceedings of the th ACM Conference on Web Science. ACM, 2019, pp. 105–114.
[28] N. Hamdi,"A guide to avoiding hate speech",The Egyptian Media Development Program,, Tech. Rep,2017
[29] R.AbuJuma,"The dictionary of hatred and hatred in the Tunisian media",Monitoring Media For Group,2013
[30] Iraqi Media House,"Hate Dictionary", Tech. Rep.,2017
[31] Syrian Center for Media and Freedom of Expression,"Study hate speech and incitement to violence in the Syrian media",Tech. Rep.,2017
[32] K. Kenyon-Dean, E. Ahmed, S. Fujimoto, J. Georges-Filteau,C. Glasz, B. Kaur, A. Lalande, S. Bhanderi, R. Belfer, N. Kanagasabai et al., “Sentiment analysis: It’s complicated!” in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume (Long Papers), 2018, pp. 1886–1895.
[33] A. Pasha, M. Al-Badrashiny, M. T. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth, “Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic.” in LREC, vol. 14, 2014, pp. 1094–1101.
[34] I. A. Al-Sughaiyer and I. A. Al-Kharashi, “Arabic morphological analysis techniques: A comprehensive survey,” Journal of the American Society for Information Science and Technology, vol. 55, no. 3, pp. 189–213, 2004.
[35] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. Bebah, and M. Shoul, “Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts,” in International Arab conference on information technology. Benghazi Libya, 2010, pp. 1–6.
[36] H. K. Aldayel and A. M. Azmi, “Arabic tweets sentiment analysis–a hybrid scheme,” Journal of Information Science, vol. 42, no. 6, pp. 782–797, 2016.
[37] S. Khoja and R. Garside, “Stemming arabic text,” Lancaster, UK, Computing Department, Lancaster University, 1999.
[38] L. S. Larkey, L. Ballesteros, and M. E. Connell, “Improving stemming for arabic information retrieval: light stemming and co-occurrence analysis,” in Proceedings of the th annual international ACM SIGIR conference on Research and development in information retrieval, 2002, pp. 275–282.
[39] V. Gupta and G. S. Lehal, “A survey of text summarization extractive techniques,” Journal of emerging technologies in web intelligence, vol. 2, no. 3, pp. 258–268, 2010.
[40] L.-P. Jing, H.-K. Huang, and H.-B. Shi, “Improved feature selection approach tfidf in text mining,” in Machine Learning and Cybernetics, . Proceedings. International Conference on, vol. 2. IEEE, 2002, pp. 944–946.
[41] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.
[42] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, “Tackling the poor assumptions of naive bayes text classifiers,” in Proceedings of the th international conference on machine learning (ICML-), 2003, pp. 616–623.
[43] J. R. Quinlan, C . : programs for machine learning. Elsevier, 2014.

[44] C. Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[45] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” Journal of machine learning research, vol. 12, no. Aug, pp. 2493–2537, 2011.

[46] K. Kawakami, “Supervised sequence labelling with recurrent neural networks,” Ph. D. dissertation, PhD thesis. Ph. D. thesis, 2008.

[47] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the Association for Computational Linguistics, vol. 5, pp. 135– 146, 2017.

[48] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv: . , 2013.

[49] A. M. Alayba, V. Palade, M. England, and R. Iqbal, “A combined cnn and lstm model for arabic sentiment analysis,” in International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Springer, 2018, pp. 179–191.

[50] B. Shin, T. Lee, and J. D. Choi, “Lexicon integrated CNN models with attention for sentiment analysis,” in Proceedings of the th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017.

[51] M. Schmitt, S. Steinheber, K. Schreiber, and B. Roth, “Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018.