Detecting Hate Speech Against Athletes in Social Media

Dana Alsagheer*, Hadi Mansourifar*, Mohammad Mahdi Dehshibi‡, Weidong Shi*
*Computer Science Department, University of Houston, Houston, USA
E-mail: dralsagh@central.uh.edu, hmansourifar@uh.edu, wshi3@central.uh.edu
‡Department of Computer Science, Universidad Carlos III de Madrid, Madrid, Spain
E-mail: mohammad.dehshibi@yahoo.com

2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA) | 978-1-6654-9960-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/IDSTA55301.2022.9923132

Abstract—**When English clubs and the game's governing bodies and organizations turned off their Facebook, Twitter, and Instagram accounts from April 30 to May 1, 2021, the fight against online racism regained a new momentum. However, the Tokyo Olympics revealed new aspects of online bullying that athletes may face during major sporting events. Despite the significant effort put into online hate speech detection research in general, hate speech detection against athletes requires a separate investigation. We show in this paper that abusive language directed at athletes is more varied and difficult to detect. We began with the introduction of the collected data from online comments aimed at three athletes competing in the Tokyo Olympics 2020. Followed by conducting an extensive classification experiments of the collected data to demonstrate its diversity in comparison to other hate speech datasets. This was done to demonstrate that Active Learning outperforms Supervised Learning in hate speech detection against athletes.**

Index Terms—**Hate Speech Detection, Active Learning, Social Media**

I. INTRODUCTION

"Social media is now sadly a regular vessel for toxic abuse. Hate has become depressingly normalised," said Kick It Out chairman Sanjay Bhandari [1]. It means that professional athletes are constant targets of online hate speech and face increased pressure as a result of social media focusing on a single individual following a loss or a specific incident [16]. "For every positive and genuine example of direct human interaction between athletes, fans and media on social media, there is also a barrage of abuse suffered by nonwhite, nonmale and nonstraight athletes whenever they underperform or are considered to have spoken out of line" [2]. Unfortunately, being spotted individually by millions of people in social media makes the professional athletes a fragile minority exposed to psychological consequences. Tokyo Olympics 2020 gave this drama a new turn which started with Simone Biles [3], continued with Tahani Alqahtani [4] a Saudi judoka and ended with Laurel Hubbard [5] a New Zealand weightlifter. In this paper, we introduce a new hate speech dataset collected from comments in social media targeting the mentioned Olympians. Although the hate speech encompass a wide range of definitions, we managed to collect certain comments as labelling justification which judged other comments as hateful. The characteristic which can distinguish our dataset from old hate speech datasets. Table 1 shows some of the justification comments. Although the comments which target each of the mentioned Olympians are categorized as hate speech, our experiments show that, they are significantly different in terms of Google Perspective Scores [6]. Besides, we train different classifiers on each subgroups using four different text vectors including Bag of Words [36], Word2Vec [24], Perspective Scores [6] and BERT embeddings [29] [8]. Finally, we use active learning [7] to train classifiers on each subgroup to show a significant improve in classification results. Here is a list of questions we answer in this paper:

A. What are the challenges for processing the collected dataset?

Collected subgroups are significantly different in terms of data distribution and instance number. It makes the training process very challenging on total data.

B. Why active learning is a better solution to classify the collected data?

The highly imbalanced nature of classes with significantly different subgroups make the active learning a better choice comparing to supervised learning since the most informative labelled instances are selected in active learning rather than all labelled instances.

Our contributions are as follows.

- We publish a new hate speech dataset collected from comments targeting the athletes in social media.
- Our experimental results show that Perspective Scores can outperform the BERT embedding features in highly skewed classes.
- We show that active learning significantly improve the hate speech detection results.

The rest of the paper is organized as follows. Section 2 reviews previous studies. Section 3 background. Section 4 demonstrate the proposed method, and section 5 and 6 provide the experimental results. Finally, section 7 provides a conclusion for the paper.

II. RELATED WORK

Hate speech detection as a general problem has been investigated before in the research community. To the best of our knowledge, no research in the data mining community

TABLE I
HATE SPEECH EVIDENCES : COMMENTS EXPRESSED BY NORMAL USERS JUDGING THE REST OF COMMENTS AS HATE SPEECH.

Ex.	Evidence
1	*The haters have arrived! The majority of you can't even do a handstand against the wall!!!* *Lay Off Her - You could never! Congrats Simone!*
2	*I thought you poor, pitiful, miserable animals in the comments were so pro USA?* *So why would you bash, harass, and bully the STAR athlete?* *Bet I know why, DRUM ROLL PLEASE..Because she's BLACK !!!* *Racist people are so miserable...Smh*
3	*These haters in the comments are acting so bogus. they act like they can do what simone does,* *they wish. they have no right to be talking smh.* *go outside and touch some grass! at least she made it on the leaderboard and for that i'm proud of her.* *she should always have her city behind her*
4	*The comments are awfully thick of transphobia and ignorance today*

TABLE II
CLASSIFICATION OF PREVIOUS WORKS IN HATE SPEECH DETECTION.

Research	Approach	Advantage	Disadvantage
[11]	*Lexicon-based*	*Simplicity*	*High false positive rate*
[21]	*Rule-learning*	*Easy to implement*	*Context sensitive*
[35]	*Non-linguistic*	*Efficiency*	*Unavailable data*
[18]	*Sentiment analysis*	*Robust*	*Needs huge data*
[34]	*Language model*	*Efficiency*	*Limited to stereotypes*
[30]	*Word Embeddings*	*State of the art*	*May need fine-tuning*
[26]	*Perspective Scores*	*Generalizable*	*The engine is not free*

TABLE III
SPECIFICATION OF COLLECTED DATA AND EACH SUBGROUP.

Subgroup	Positive (%)	Negative (%)	Total Instances Number
Simone	72	28	120
Tahani	32	68	126
Laurel	70	30	40
Total	57	43	286

has studied this problem with a special focus on a particular minority group like professional athletes. The only published research in this area comes from Public Health Domain [32]. In this research, Vveinhardt et al. analyzed consequences of bullying in the contexts of athletes' emotional state and career. In this section, we review some of the previous works in general hate speech detection. Table 2 summarizes the hate speech detection methods with a brief advantage and disadvantage column as explanation. Most of the previous hate speech detection researches are based on a dataset. That's why we briefly introduce some of the major hate speech datasets as tabulated in Table 4. Note that, WARNER dataset [34] is not publicly available and it has been manually annotated into seven categories including anti-semitic, anti-black, anti-Asian, anti-woman, anti-Muslim, anti-immigrant or other hate(anti-gay and anti-white).

TABLE IV
LIST OF MAJOR HATE SPEECH DATASETS AND THEIR SPECIFICATIONS.

Dataset	Size	Pos (%)	Neg (%)	Source
Davidson [15]	*24,802*	*5.77*	*94.23*	*Twitter*
WARNER [34]	*9000*	-	-	*Yahoo*
DJURIC [18]	*209776*	*26*	*74*	*Yahoo*
QIAN [28]	*22,324*	*23.5*	*76.5*	*Reddit*
BENIKOVA [10]	*36*	*33*	*67*	*Twitter*
Clubhouse [26]	*468*	*26*	*74*	*Clubhouse*
TweetBLM [25]	*9165*	*33*	*67*	*Twitter*

III. DATASET COLLECTION, ANNOTATION AND ANALYSIS

In this section, we present a detailed description of data collection from motivation to data analysis.

A. Tokyo Games Story

On July 28, 2021, in the midst of the Tokyo Olympics, a breaking news shocked the world of sports: "Gymnastics superstar and defending Olympic champion Simone Biles has withdrawn from Thursday's individual all-around competition at the Tokyo Games to focus on her mental well-being" [3]. This incident gave the haters an excuse to barrage social media with abusive comments, targeting Simone Biles. Two days later, when the waves of cyber bullying [20] [14] against Simone Biles seemed to be subsided, the next round of online attacks started against another athlete, inflaming the racism on social media again: "A Saudi Arabian judoka faced off against her Israeli opponent at the Tokyo Olympics today - defying pressure to follow the lead of two other Muslim athletes who boycotted their bouts. Tahani al-Qahtani fought against Raz

TABLE V
SAMPLE HATE SPEECH INSTANCES IN COLLECTED DATASET CATEGORIZED IN THREE SUBGROUPS.

Ex.	Subgroup	Hate Speech Instance
1	Simone	*She should smoke some weed to calm her mind and be mentality relaxed with all the weight of the world on her shoulders.*
2	Simone	*That's nice... I'm glad she can realize that she' s a human being! she should realize she's a Selfish person as well and that would just be icing in the cake*
3	Simone	*Biles is no GOAT, she is a COWard*
4	Tahani	*My message to Tahani Al-Qahtani.lose weight. Pray rather than play. Pray and ask forgiveness and stay in the kitchen. Take care of your studies. Leave sports to its people. The entire Saudi players are failures and embarrassed us with a shameful appearance..*
5	Tahani	*God humiliated her and she humiliated his country*
6	Tahani	*A scandal.normalization..a loss of employment.In addition to all of this, it took severe beatings and a humiliating loss. you deserve it*
7	Laurel	*So it's a beta male competing against woman because he is to weak to compete against men.*
8	Laurel	*That's a Larry not Laurel... the worlds gone mad*
9	Laurel	*I absolutely will not watch anything. Between this thief stealing the gold from the women and the women's soccer team being a disgrace to our country, what's the point.*
10	Simone	*Good thing she's not a mom because you don't get to quit for a mental health break.*
11	Simone	She is a Quitter. She failed her team, her city, her state, and her country. She choked. She will go down in history as one of the worst quitters ever.

Hershko in the women's 78kg category at Tokyo on Friday, before the pair clasped hands and raised them in the air as a show of solitary when the bout was over" [4]. Unfortunately, the dark side of Tokyo games got a new turn when Laurel Hubbard a New Zealand weightlifter started her competitions as the first openly transgender woman [5]. The online attacks in this last case were so harsh that normal comments were extremely rare. Perhaps, for the first time in the history of data mining, collected data in the third subgroup belonging to Laurel Hubbard skewed toward positive (hateful) instances.

B. Data Collection

In order to gather a diverse dataset, we collected data from three different platforms namely Facebook, youtube, and Twitter during four weeks. Our method was manually select videos that contain acts of hate speech against one of the athletes from youtube and collect all comments. On Facebook, we use the official page of all three athletes and collect all the comments negative or positive comments. Twitter was a hashtag-based collected data method. Usually, data from social media is hard to analyze because it contains a lot of grammar mistakes, syntactic errors, and a lot of ad-hoc spelling. Preparing the data to feed into the classifier is crucial. For instance, we lower cased the data, removed Email addresses, corrupted, incorrectly formatted, duplicated, or incomplete data within a dataset. The annotation was done by four different individuals since automatic labeling is likely to be noisy. Table 3 summarizes the number of instances in each subgroup.

Here is some facts about each subgroups:

- In the Simone subgroup, negative instances (normal comments) are very similar to each other. We decided to keep the redundancy low in our dataset to avoid over-fitting in our trained models.
- In the Laurel subgroup, normal instances were very rare and many redundant hate speech instances were removed later from the initial data.
- In the Tahani subgroup, three different annotators provided the labels with 100 percent agreement.
- Instances in the Simone subgroup are more diverse with a more polite tone. That is why we selected more sample instances from this subgroup to be shown in Table 5.

Some of the hate speech instances belonging to each subgroup are tabulated in Table 5. Figure 1 illustrates the distribution of each subgroup using toxicity scores and T-SNE method [33].

IV. ACTIVE LEARNING FOR HATE SPEECH DETECTION

Active learning [7] is known as a technique to address the unlabeled training instance in which the learner tries to obtain the most informative examples via asking a limited number of labels from user. Yet, in case of available labels, active learning can still be useful to obtain the most informative instances from the training set. One of the practiced applications of active learning is imbalanced classification where the learner focuses to select the most representative instances to minimize the impact of skewed regions. To the best of our knowledge, no effort has been made to take advantage of active learning in hate speech detection. In the following section, we demonstrate the sampling strategy in which the learner tries to find the most informative examples. Figure 2 shows Active Learning diagram.

V. EXPERIMENTAL SETUPS

In this section, we review the detail settings of our experiments including base classifiers, feature extraction and performance measures.

A. Feature Extraction

We used three different feature extraction methods as follows.

- Bag of words [23], [36]: The 'bag of words' is the word vector for each instance in the dataset which is a simple word count for each instance where each position of the vector represents a word. We used ngram range of (1,2) which means that each vector is divided into single words and also pairs of consecutive words. We also set the max size of the word vector to 10,000 words.
- Word2vec : Word2Vec [24], [27] is one of the most popular techniques to learn word embeddings using shallow

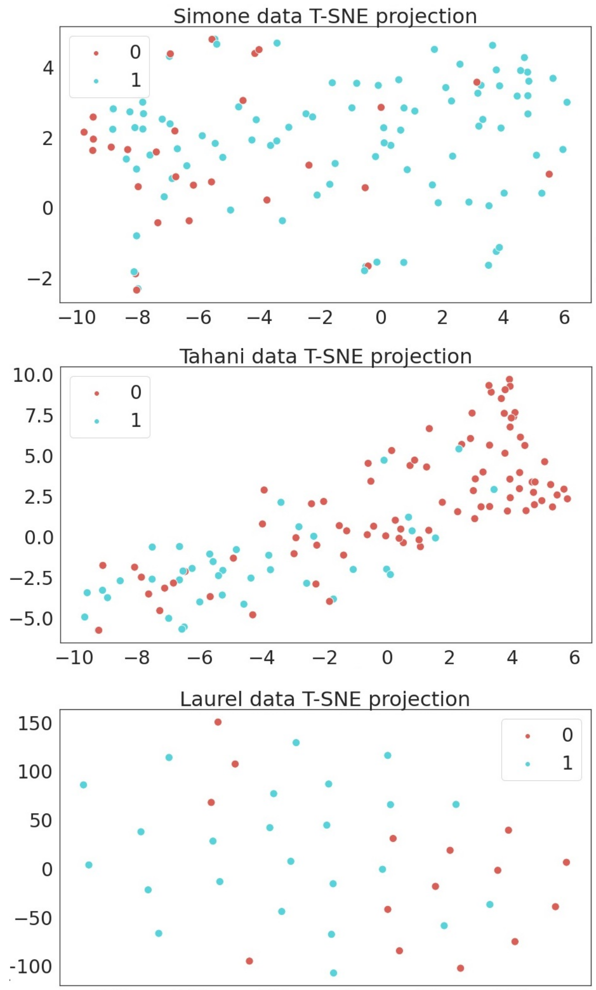

Fig. 1. Perspective data distribution in 2D using T-SNE [33]: Data distribution is significantly different in each subgroup. (0s=normal instances, 1s= hate instances)

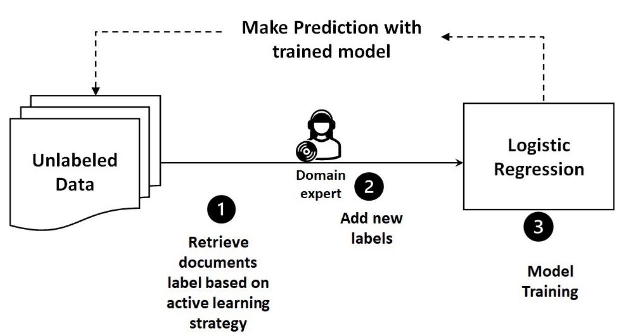

Fig. 2. Illustration of the active learning approach [12].

neural network. It was developed by Tomas Mikolov in 2013 at Google. As the pre-trained model we used Google News corpus trained on 3 million words and phrases. this model provides 300-dimensional vectors as the transferred data.

- Perspective Scores: They are high level features calculated by different trained classifiers. We passed all the records to Google Perspective API and collected 9 Perspective Scores per input vector as mentioned in previous section. We applied based classifiers without over-sampling [9], [13], [22] on the transferred vectors.
- BERT features: Bidirectional Encoder Representations from Transformers (BERT) [17] is a transformer-based machine learning technique for natural language processing developed by Google. We pass the input sequence of tokens to the BERT pre-trained model to extract 768 contextualized features. .

B. Perspective Scores

In order to extract high level features from hate speech instances, we use Perspective API [6]. Perspective API was developed by Jigsaw and Google's Counter Abuse Technology team as a part of the Conversation-AI project. The API provides several pre-trained models to compute several scores between 0 and 1 for different categories as follows [19].

- toxicity is a "rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion."
- severe toxicity is a "very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective."
- identity attack are "negative or hateful comments targeting someone because of their identity."
- insult is an "insulting, inflammatory, or negative comment towards a person or a group of people."
- profanity are "swear words, curse words, or other obscene or profane language"
- threat "describes an intention to inflict pain, injury, or violence against an individual or group."

All the trained models use Convolutional Neural Networks (CNNs), trained with GloVe word embeddings [27] and fine-tuned during training on data from online sources such as Wikipedia and The New York Times [19].

C. Evaluation measures

In this section, we present the evaluation measures used in our experiments.

1) Classification measures: Classifier performance metrics are typically evaluated as follows. TP (True Positive) is the number of correctly classified positive instances. FN (False Negative) is the number of incorrectly classified positive instances. FP (False Positive) is the number of incorrectly classified negative instances. TN (True Negative) is the number of correctly classified negative instances. The three performance measures including precision, recall and F1 are defined by following formulae.

Recall = TP/(TP+ FN),

TABLE VI
ACTIVE LEARNING PARAMETER SETTING.

	Simone	Tahani	Laurel	Total
Initial Size	50	50	20	90
Queries	30	30	5	30

Precision = TP/(TP+ FP),
F1 = (2* Recall * Precision) /(Recall+ Precision)

D. Base Classifier

After extracting mentioned text features, we used Logistic Regression (LR) [31] as base classifier for hate speech detection. LR uses logistic function and log odds to perform a binary classification.

E. Implementation

To implement the feature extraction methods and base classifiers, we used python libraries including Sklearn, Pandas, etc. All the codes and three versions of labelled data would be publicly available after paper publication.

F. Active Learning Setup

To run the active learning experiments we need to set two parameters including number of initial instances and number of queries. Table 6 shows the values for each parameter in our experiments. Note that, we used uncertainty sampling as sampling strategy in all the experiments. Figure 3 shows Active Learning training accuracy per each query.

VI. Experimental Results

In this section, we present the classification results and discussions.

A. Supervised Learning Results

The supervised learning results are shown in Table 7 and can be summarized as follows.

- Word2Vec obtains the best results in each subgroup in terms of accuracy, precision and f1 score.
- Bag of Words obtains the best results in terms of recall in all the supervised learning experiments.
- Perspective Score obtains the best results in total data in terms of accuracy, precision and f1 score.

1) Why Perspective Scores are superior in total data?: In general, high level features show their discriminative power in complex boundaries. However, in less complicated data like each subgroup they cause over-fitting. That's why Perspective scores does not perform well on each subgroup but outperform the rest of feature extraction methods in total data.

B. Active Learning Results

The Active learning results can be summarized as follows.

- Active learning outperforms supervised learning in terms of f1 score as shown in Table 8 and Figure 4.
- Perspective Scores obtain the best f1 score on total data in active learning experiments.

TABLE VII
SUPERVISED LEARNING RESULTS ON SUBGROUPS AND TOTAL DATA. THE BEST RESULTS IN EACH GROUP ARE HIGHLIGHTED IN BOLD.

	Classification Results on Simone Data			
	Accuracy	*Precision*	*Recall*	*F1 score*
BOW	0.8	0.813	**0.96**	0.8714
W2V	**0.86**	**0.886**	0.9488	**0.9113**
BERT	0.825	0.8612	0.926	0.8856
PS	0.7643	0.7643	1.0	0.8636
	Classification Results on Tahani Data			
BOW	0.7134	0.57	0.3933	0.4392
W2V	**0.7224**	**0.6583**	0.5521	**0.5421**
BERT	0.6891	0.5747	**0.5526**	0.5112
PS	0.7096	0.616	0.385	0.453
	Classification Results on Laurel Data			
BOW	0.6	0.6	**0.8333**	0.6704
W2V	**0.8**	**0.7833**	0.7916	**0.7757**
BERT	0.675	0.6166	0.7	0.6428
PS	0.6166	0.7333	0.65	0.6633
	Classification Results on Total Data			
BOW	0.696	0.6944	**0.8180**	0.7409
W2V	0.7039	0.7329	0.7305	0.7257
BERT	0.6786	0.6989	0.7265	07023
PS	**0.7438**	**0.7738**	0.7339	**0.7472**

TABLE VIII
ACTIVE LEARNING RESULTS ON SUBGROUPS AND TOTAL DATA.

	Active Learning Results on Simone Data			
	Accuracy	*Precision*	*Recall*	*F1 score*
PS	0.8474	0.875	0.9333	0.9032
BERT	**0.8666**	0.8584	**0.9891**	**0.9191**
BOW	0.8666	**0.88**	0.9565	0.9166
W2V	0.825	0.8198	0.9891	0.8965
	Active Learning Results on Tahani Data			
PS	**0.792**	**0.666**	0.7317	**0.6976**
BERT	0.7380	0.6363	0.5	0.56
BOW	0.6587	0.4933	**0.8809**	0.6324
W2V	0.7063	0.5396	0.8095	0.6496
	Active Learning Results on Laurel Data			
PS	**0.9230**	**0.9565**	0.916	**0.9361**
BERT	0.8	0.7575	**1.0**	0.862
BOW	0.85	0.8064	1.0	0.8928
W2V	0.85	0.8064	1.0	0.8928
	Active Learning Results on Total Data			
PS	**0.8014**	**0.7734**	0.9032	**0.8333**
BERT	0.7167	0.6681	**0.9748**	0.7928
BOW	0.7307	0.6798	0.9748	0.8010
W2V	0.6258	0.6	0.9748	0.7434

1) Why Active Learning outperforms the supervised learning in collected data?: Active learning can alleviate the impact of skewed classes by selecting the most informative instances among the unseen training data. Besides, active learning minimizes the possibility of wrong labels in training data via giving the domain expert a second chance to judge the instances at the run time.

VII. Conclusion

We introduced a new hate speech dataset with a special focus on professional athletes as one of the prominent targets

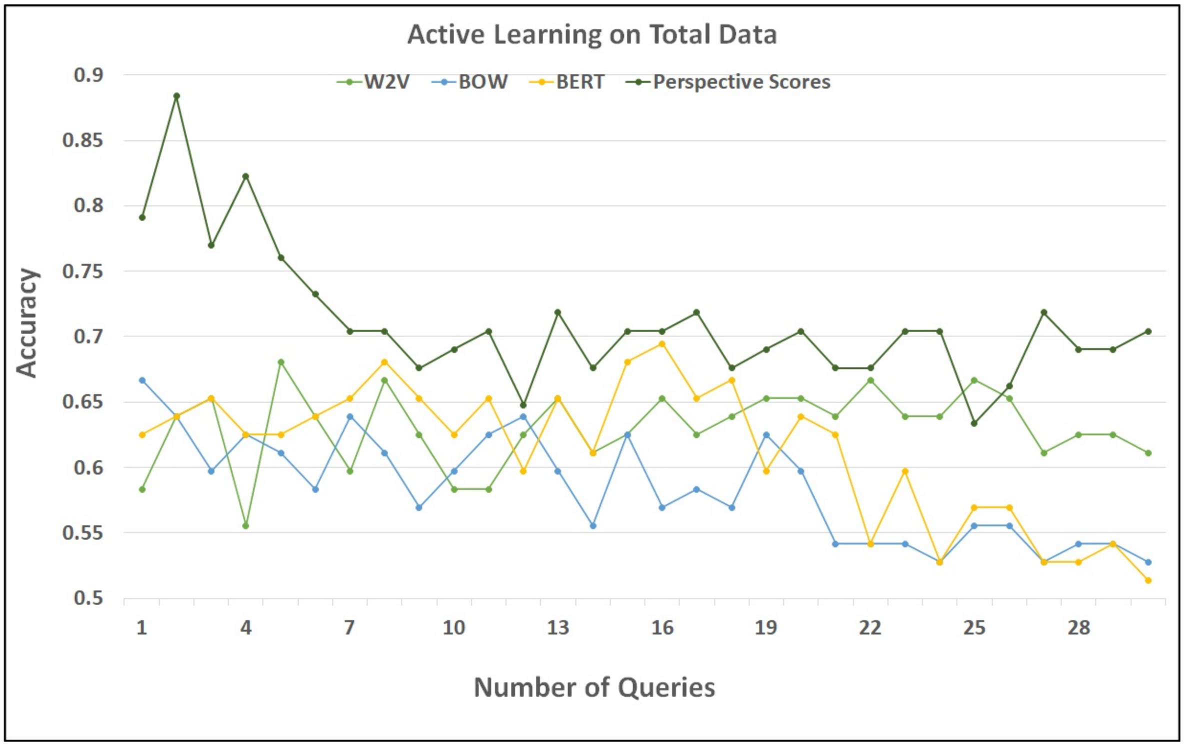

Fig. 3. Active learning training accuracy per each query.

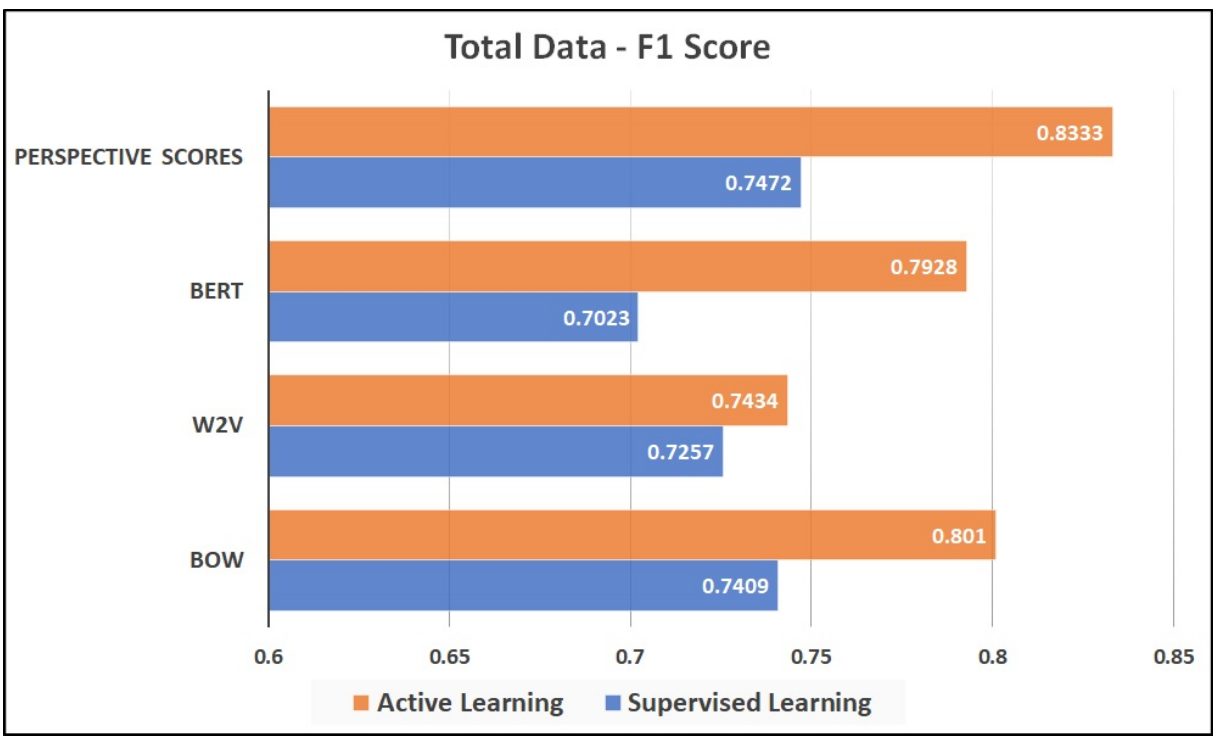

Fig. 4. Active learning versus supervised learning.

of abusing language in social media. We demonstrated that, hate speech against athletes is significantly different from already known hate speech instances. We tested different learning methodologies including supervised learning and active learning. Our experiments showed that, active learning and Perspective Scores can be more successful in case of highly skewed classes toward positive instances comparing to supervised learning and other feature extraction methods.

REFERENCES

[1] https://www.cnn.com/2021/04/25/football/english-football-social-media-boycott-spt-intl/index.html, 2021.

[2] https://www.dw.com/en/naomi-osaka-a-victim-of-broken-relationship-between-athletes-and-media/a-57741920, 2021.

[3] https://timesofoman.com/article/104710-simone-biles-withdraws-from-olympic-all-around-competition-to-focus-on-mental-health, 2021.

[4] https://www.dailymail.co.uk/news/article-9843705/Saudi-judo-Olympian-ignores-pressure-boycott-bout-against-Israeli.html, 2021.

[5] https://www.nytimes.com/2021/07/31/sports/laurel-hubbard-trans-weight-lifting.html, 2021.

[6] https://www.perspectiveapi.com/, 2021.

[7] Umang Aggarwal, Adrian Popescu, and Céline Hudelot. Active learning for imbalanced datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1428–1437, 2020.

[8] Walaa Alnasser, Ghazaleh Beigi, and Huan Liu. Privacy preserving text representation learning using bert. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 91–100. Springer, 2021.

[9] Sukarna Barua, Md Monirul Islam, Xin Yao, and Kazuyuki Murase. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering*, 26(2):405–425, 2012.

[10] Darina Benikova, Michael Wojatzki, and Torsten Zesch. What does this imply? examining the impact of implicitness on the perception of hate speech. In *International Conference of the German Society for Computational Linguistics and Language Technology*, pages 171–179. Springer, 2017.

[11] Peter Burnap and Matthew Leighton Williams. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. 2014.

[12] Andres Carvallo, Denis Parra, Hans Lobel, and Alvaro Soto. Automatic document screening of medical literature using word and text embeddings in an active learning setting. *Scientometrics*, 125(3):3047–3084, 2020.

[13] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[14] Lu Cheng, Ahmadreza Mosallanezhad, Yasin Silva, Deborah Hall, and Huan Liu. Mitigating bias in session-based cyberbullying detection: A non-compromising approach. In *Proceedings of ACL*, 2021.

[15] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[16] Mohammad Mahdi Dehshibi, Bita Baiani, Gerard Pons, and David Masip. A deep multimodal learning approach to perceive basic needs of humans from instagram profile. *IEEE Transactions on Affective Computing*, 2021.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30, 2015.

[19] Paula Fortuna, Juan Soler, and Leo Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794, 2020.

[20] Suyu Ge, Lu Cheng, and Huan Liu. Improving cyberbullying detection with user interaction. In *Proceedings of the Web Conference 2021*, pages 496–506, 2021.

[21] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[22] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[23] Haiyun Jiang, Yanghua Xiao, and Wei Wang. Explaining a bag of words with hierarchical conceptual labels. *World Wide Web*, pages 1–21, 2020.

[24] Suhyeon Kim, Haecheong Park, and Junghye Lee. Word2vec-based latent semantic analysis (w2v-lsa) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications*, 152:113401, 2020.

[25] Sumit Kumar, Raj Ratn Pranesh, and Subhash Chandra Pandey. Tweetblm: A hate speech dataset and analysis of black lives matter-related microblogs on twitter.

[26] Hadi Mansourifar, Dana Alsagheer, Reza Fathi, Weidong Shi, Lan Ni, and Yan Huang. Hate speech detection in clubhouse. *arXiv preprint arXiv:2106.13238*, 2021.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[28] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*, 2019.

[29] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[30] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, 2017.

[31] Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, 5(1):1–16, 2020.

[32] Jolita Vveinhardt, Vilija Bite Fominiene, and Regina Andriukaitiene. Encounter with bullying in sport and its consequences for youth: Amateur athletes' approach. *International Journal of Environmental Research and Public Health*, 16(23), 2019.

[33] Vincent D Warmerdam, Thomas Kober, and Rachael Tatman. Going beyond t-sne: Exposing\texttt {whatlies} in text embeddings. *arXiv preprint arXiv:2009.02113*, 2020.

[34] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26, 2012.

[35] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[36] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.