Educational Administration: Theory and Practice
2024, 30(5), 12663-12668
ISSN:2148-2403
https://kuey.net/ **Research Article**

Automatic Hate Speech Detection and the hassle of Offensive Language

Puspendu Biswas[1*] Donavalli Haritha[2]

[1*]Ph.D. Scholar, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, puspendu.biswas82@gmail.com
[2]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, haritha_donavalli@kluniversity.in

Citation: Puspendu Biswas, et al (2024), Automatic Hate Speech Detection and the hassle of Offensive Language , *Educational Administration: Theory and Practice*,12663-12668, DOI: 10.53555/kuey.v30i5.4005

ARTICLE INFO **ABSTRACT**

A key task for automatic hate-speech detection on social media is the separation of hate speech from different instances of offensive language. Lexical detection strategies tend to have low precision due to the fact they classify all messages containing precise terms as hate speech and previous work the use of supervised gaining knowledge of has failed to differentiate among the two classes. We used a crowd-sourced hate speech lexicon to acquire tweets containing hate speech keywords. We use crowdsourcing to label a pattern of those tweets into three classes: those containing hate speech, only offensive language, and those with neither. We educate a multi-magnificence classifier to distinguish among those one-of-a-kind categories. near analysis of the predictions and the errors suggests when we can reliably separate hate speech from different offensive language and while this differentiation is extra difficult. we discover that racist and homophobic tweets are much more likely to be categorized as hate speech but that sexist tweets are normally labeled as offensive. Tweets without specific hate key phrases also are more difficult to categories.

INDEX TERMS Hate Speech · Dataset Presentation · Machine Learning · Binary/ Multi-label Classification · Active Learning

I. INTRODUCTION

What hate speech and while does it fluctuate from offensive language? No formal definition exists but there is a consensus that it's miles speech that targets disadvantaged social companies in a way this is doubtlessly harmful to them (Jacobs and Potter 2000; Walker 1994). within the use, hate speech is covered below the free speech provisions of the first change, but it has been extensively debated inside the legal sphere and almost about speech codes on university campuses. In many countries, along with the UK, Canada, and France, there are legal guidelines prohibiting hate speech, which tends to be defined as speech that goals minority agencies in a way that could promote violence or social disease. people convicted of the use of hate speech can frequently face large fines or even imprisonment. these laws increase to the internet and social media, main many sites to create their own provisions towards hate speech. each fb and Twitter have answered to grievance for no longer doing enough to prevent hate speech on their sites via instituting rules to limit the usage of their platforms for attacks on people based totally on characteristics like race, ethnicity, gender, and sexual orientation, or threats of violence toward others.1 Drawing upon these definitions, we outline hate speech as language this is used to expresses hatred closer to a focused organization or is supposed to be derogatory, to humiliate, or to insult the individuals of the group. In intense cases this may additionally be language that threatens or incites violence, however limiting our definition best to such instances might exclude a large percentage of hate speech. Importantly, our definition does now not include all instances of offensive language because people frequently use terms which might be particularly offensive to certain corporations but in a qualitatively one-of-a-kind manner. for example, a few African individuals regularly use the term n*gga2 in everyday language online (Warner and Hirschberg 2012), human beings use terms like h*e and b*tch while quoting rap lyrics, and teens use homophobic slurs like f*g as they play video games. Such language is general on social media (Wang et al. 2014), making this boundary condition essential for any usable hate speech detection system. preceding paintings on hate speech detection has diagnosed this trouble but much research nevertheless tend to conflate hate speech and offensive language. on this paper we label tweets into three categories: hate speech, offensive language, or neither. We teach a version to differentiate among those classes and then examine the effects so as to better recognize how we can distinguish between them. Our effects display that satisfactory-grained labels can assist within the venture of hate speech detection and highlights some of the important thing demanding situations to accurate class. We conclude that future paintings should higher account for context and the heterogeneity in hate speech utilization.

2. RELATED WORK

Bag-of-words strategies tend to have high don't forget but cause high quotes of fake positives since the presence of offensive words can cause the misclassification of tweets as hate speech (Kwok and Wang 2013; Burnap and Williams 2015). focusing on anti-black racism, Kwok and Wang find that 86% of the time the motive a tweet changed into classified as racist became as it contained offensive words. Given the especially excessive prevalence of offensive language and “curse words” on social media this makes hate speech etection particularly difficult (Wang et al. 2014). The difference between hate speech and other offensive language is regularly primarily based upon diffused linguistic distinctions, for instance tweets containing the word n*gger are more likely to be categorized as hate speech than n*gga (Kwok and Wang 2013). Many can be ambiguous, for instance the phrase homosexual can be used both pejoratively and in different contexts unrelated to hate speech (Wang et al. 2014). Syntactic functions have been leveraged to better identify the objectives and depth of hate speech, for example sentences wherein a applicable noun and verb arise (e.g. kill and Jews) (Gitari et al. 2015), the POS trigram “DT jewish NN” (Warner and Hirschberg 2012), and the syntactic shape I , e.g. “I f*cking hate white people” (Silva et al. 2016). other supervised approaches to hate speech type have unluckily conflated hate speech with offensive language, making it difficult to ascertain the extent to which they're truly identifying hate speech (Burnap and Williams 2015; Waseem and Hovy 2016). Neural language fashions display promise within the assignment but current paintings has used training records has a in addition vast definition of hate speech (Djuric et al. 2015). Non-linguistic capabilities just like the gender or ethnicity of the author can assist enhance hate speech class but this data is frequently unavailable or unreliable on social media (Waseem and Hovy 2016).

3. DATA

We begin with a hate speech lexicon containing phrases and phrases recognized via internet customers as hate speech, compiled by Hatebase.org. the use of the Twitter API we searched for tweets containing terms from the lexicon, ensuing in a pattern of tweets from 33,458 Twitter customers. We extracted the time-line for each consumer, resulting in a hard and fast of 85.4 million tweets. From this corpus we then took a random sampleof 25k tweets containing terms from the lexicon and had them manually coded through CrowdFlower (CF) people.employees have been requested to label each tweet as one in all 3 categories: hate speech, offensive but not hate speech, or neither offensive nor hate speech.

They have been supplied with our definition at the side of a paragraph explaining it in further element. users have been asked to suppose not just about the words acting in a given tweet but about the context wherein they have been used. They had been instructed that the presence of a specific word, however offensive, did now not always suggest a tweet is hate speech

every tweet changed into coded with the aid of three or greater people.The intercoder-agreement rating supplied through CF is 92%. We use the general public decision for every tweet to assign a label. some tweets have been no longer assigned labels as there was no majority elegance. This effects in a sample of 24,802 categorized tweets. simplest five% of tweets were coded as hate speech by means of the majority of coders and best 1.3% have been coded unanimously, demonstrating the imprecision of the Hatebase lexicon. this is an awful lot lower than a comparable have a look at the usage of Twitter, where 11.6% of tweets were flagged as hate speech (Burnap and Williams 2015), in all likelihood due to the fact we use a stricter criteria for hate speech. the majority of the tweets have been considered to be offensive language (76% at 2/three, fifty three% at 3/3) and the remainder had been considered to be non-offensive (16.6% at 2/3, eleven.8% at three/3). We then built features from these tweets and used them to educate a classifier. capabilities We lowercased each tweet and stemmed it the use of the Porter stemmer,3 then create bigram, unigram, and trigram capabilities, every weighted through its TF-IDF. To seize facts about the syntactic shape we use NLTK (fowl, Loper, and Klein 2009) to assemble Penn element-of-Speech (POS) tag unigrams, bigrams, and trigrams.

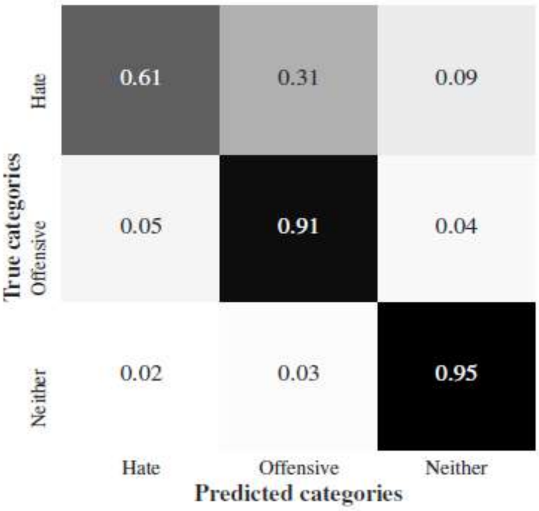
To seize the great of each tweet we use changed Flesch-Kincaid Grade stage and Flesch studying Ease scores, in which the quantity of sentences is fixed at one. We additionally use a sentiment lexicon designed forsocial media to assign sentiment scores to each tweet (Hutto and Gilbert 2014). We also include binary and matter indicators for hashtags, mentions, retweets, and URLs, as well as functions for the quantity of characters, words, and syllables in each tweet.

4. Model

We first use a logistic regression with L1 regularization to reduce the dimensionality of the facts. We then take a look at a diffusion of models which have been used in prior work: logistic regression, na¨ıve Bayes, selection bushes, random forests, and linear SVMs. We tested each model using five-fold cross validation, preserving out 10% of the sample for assessment to help preventover-becoming. After the use of a grid-seek to iterate over the fashions and parameters we find that the Logistic Regression and Linear SVM tended to carry out drastically better than different models. We determined to use a logistic regression with L2 regularization for the final version because it more without difficulty allows us to have a look at the expected possibilities of class membership and has done well in preceding papers (Burnap and Williams 2015; Waseem and Hovy 2016). We trained the very last version the usage of the complete dataset and used it to are expecting the label for every tweet. We use a one-versus-relaxation framework wherein a separate classifier is trained for each elegance and the magnificence label with the best predicted chance across all classifiers is assigned to every tweet. All modeling was performing the use of scikit-study (Pedregosa and others 2011)..

4.1 Results

The nice performing model has an usual precision 0.ninety one, do not forget of zero.ninety, and F1 rating of 0.90. searching at parent 1, but, we see that almost 40% of hate speech is misclassified: the precision and recall rankings for the hate magnificence are 0.44 and zero.sixty one respectively. most of the misclassification happens within the top triangle of this matrix, suggesting that the version is biased toward classifying tweets as less hateful or offensive than the human coders. far fewer tweets are categorized as extra offensive or hateful than their proper category; approximately 5% of offensive and a couple of% of harmless tweets have been erroneously labeled as hate speech. To discover why those tweets were misclassified we now appearance greater closely at the tweets and their expected training.



Tweets with the very best expected possibilities of being hate speech generally tend to contain multiple racial or homophobic slurs, e.g. @JuanYeez shut yo beaner ass up sp*c and hop your f*ggot ass returned throughout the border little n*gga and RT @eBeZa: silly f*cking n*gger LeBron. You flipping jungle bunny monkey f*ggot. other tweets tend to be successfully diagnosed as hate when they contained strongly racist or homophobic terms like n*gger and f*ggot. curiously, we also find instances wherein people use hate speech to reply to other hate audio system, together with this tweet in which a person uses a homophobic slur to criticize someone else’s racism: @MrMoonfrog @RacistNegro86 f*ck you, stupid ass coward b*tch f*ggot racist piece of sh*t. Turning to real hate speech categorised as offensive it seems that tweets with the very best predicted chance of being offensive are surely much less hateful and were perhaps mislabeled, for example whilst you recognize how curiosity is a b*tch #CuriosityKilledMe might also have been erroneously coded as hate speech if humans idea that interest changed into someone, and Why no boycott of racist ”redskins”? #Redskins #ChangeTheName consists of a slur but is actually in opposition to racism. it's miles likely that coders skimmed these tweets too fast, selecting out phrases or phrases that seemed to be hateful without considering the context. Turning to borderline instances, where the opportunity of being offensive is marginally better than hate speech, it appears that most of the people are hate speech, both directed toward different Twitter users, @MDreyfus @NatFascist88 Sh*t your ass your mothers p*ssy u Jew b*stard. Ur instances coming. Heil Hitler! and fashionable hateful statements like My recommendation of the day: if your a tranny...go f*ck yourself!. these tweets healthy our definition of hate speech but have been possibly misclassified due to the fact they do now not incorporate any of the phrases most strongly related to hate speech. in the end, the hateful tweets incorrectly categorized as neither have a tendency not to contain hate or curse words, for example If a few one isn’t an Anglo-Saxon Protestant, they don't have any right to be alive in the US. None at all, they may be foreign grime incorporates a terrible term, grime however no slur towards a specific organization. We additionally see that rarer varieties of hate speech, for instance this anti-chinese language statement each slant in #la should be deported. the ones scum haven't any right to be here. Chinatown ought to be bulldozed, are incorrectly categorized. at the same time as the classifier plays nicely at customary types of hate speech, especially anti-black racism and homophobia, however is less reliable at detecting types of hate speech that arise from time to time, a problem referred to via Nobata et al. (2016).

A key flaw in a lot preceding work is that offensive language is mislabeled as hate speech due to an excessively broad definition. Our multi-class framework lets in us to decrease these errors; most effective 5% of our proper offensive language become categorised as hate. The tweets correctly categorised as offensive have a tendency to contain curse phrases and regularly sexist language, e.g. Why you involved bout that other h*e? Cuz that other h*e aint involved bout another h*e and i knew Kendrick Lamar was onto something while he stated "I name a b*tch a b*tch, a h*e a h*e, a female a woman". many of these tweets incorporate sexist terms like b*tch, p*ssy, and h*e. Human coders appear to keep in mind racists or homophobic terms to be hateful however recall words which are sexist and derogatory toward girls to be simplest offensive, consistent prior findings (Waseem and Hovy 2016). searching at the tweets misclassified as hate speech we see that many incorporate multiple slurs, e.g. @SmogBaby: these h*es be lyin to all of us n*ggas and My n*gga mister meaner simply hope again inside the b*tch. at the same time as these tweets include terms that may be taken into consideration racist and sexist it's miles obvious than many Twitter customers use this sort of language in their regular communications. after they do comprise racist language they have a tendency to contain the term n*gga in preference to n*gger, in keeping with the findings of Kwok and Wang (2013). We also observed some ordinary phrases together with these h*es ain't loyal that have been definitely lyrics from rap songs that customers have been quoting. type of such tweets as hate speech leads us to overestimate the prevalence of the phenomenon. at the same time as our model nonetheless misclassifies some offensive language as hate speech we are able to avoid the extensive majority of those errors by means of differentiating among the 2. in the end, turning to the neither class, we see that tweets with the highest predicted chance of belonging to this elegance all look like harmless and have been covered within the pattern due to the fact they contained terms protected inside the Hatebase lexicon along with charlie and bird which are usually now not utilized in a hateful way. Tweets with usual wonderful sentiment and higher readability scores are much more likely to belong to this elegance. The tweets on this category that have been misclassified as hate or offensive tend to mention race, sexuality, and other social classes which can be centered by means of hate speakers. most appear like misclassifications appear to be due to on the presence of probably offensive language, as an instance He's a damn desirable actor. As a homosexual guy it's tremendous to see an brazenly queer actor given the lead function for a primary movie includes the potentially the offensive phrases homosexual and queer however uses them in a superb feel. This hassle has been encountered in preceding studies (Warner and Hirschberg 2012) and illustrates the importance of taking context under consideration. We also discovered a small wide variety of cases wherein the coders seem to have neglected hate speech that become correctly diagnosed by way of our model, e.g. @mayormcgunn @SenFeinstein White people need those weapons to protect themselves from the subhuman trash your sort unleashes on us. This locating is consistent with previous paintings that has found amateur coders to often be unreliable at identifying abusive content (Nobata et al. 2016; Waseem 2016).

4.2 Conclusions

If we conflate hate speech and offensive language then we erroneously remember many human beings to be hate audio system (mistakes within the decrease triangle of discern 1) and fail differentiate among commonplace offensive language and extreme hate speech (errors in the upper triangle of discern 1). Given the criminal and ethical implications of hate speech it's far vital that we're able to appropriately distinguish between the 2. Lexical methods are effective ways to discover potentially offensive terms however are faulty at figuring out hate speech; best a small percent of tweets flagged by the Hatebase lexicon have been taken into consideration hate speech via human coders.four whilst computerized classification methods can gain fantastically excessive accuracy at differentiating between these distinctive lessons, close evaluation of the consequences indicates that the presence or absence of particular offensive or hateful terms can each assist and preclude accurate class. steady with preceding paintings, we discover that certain phrases are especially beneficial for distinguishing between hate speech and offensive language. even as f*g, b*tch, and n*gga are used in each hate speech and offensive language, the terms f*ggot and n*gger are normally related to hate speech. among the tweets considered most hateful incorporate a couple of racial and homophobic slurs. while this allows us to without difficulty identify some of the greater egregious instances of hate speech it means that we're more likely to misclassify hate speech if it doesn't comprise any curse words or offensive terms. To extra appropriately classify such cases we should find assets of education records which might be hateful without always the use of particular key phrases or offensive language. Our outcomes additionally illustrate how hate speech can be used in 4If a lexicon have to be used we advocate that a smaller lexicon with better precision is most popular to a larger lexicon with better bear in mind.we've made a greater confined model of the Hatebase lexicon available right here: https://github.com/t-davidson/ hate-speech-and-offensive-language. distinct ways: it is able to be directly ship to a person or institution of humans centered, it is able to be espoused to no person in particular, and it could be used in communication between humans. destiny paintings have to distinguish between these one of a kind makes use of and look more closely on the social contexts and conversations in which hate speech occurs. We need to also have a look at extra closely the those who use hate speech, focusing both on their individual traits and motivations and on the social systems they are embedded in. Hate speech is a tough phenomenon to define and isn't monolithic. Our classifications of hate speech have a tendency to mirror our own subjective biases. people pick out racist and homophobic slurs as hateful however have a tendency to look sexist language as simply offensive. while our effects show that human beings perform well at identifying a number of the more egregious instances of hate speech, especially anti-black racism and homophobia, it's far essential that we're cognizant of the social biases that enter into our algorithms and destiny paintings need to purpose to perceive and accurate those biases.

References

1. Alharthi, D.N., Regan, A.C.: Social engineering defense mechanisms: A taxonomy and a survey of employees' awareness level. In: K. Arai, S. Kapoor, R. Bhatia (eds.) Intelligent Computing - Proceedings of the 2020 Computing Conference, Volume 1, SAI 2020, London, UK, 16-17 July 2020, Advances in Intelligent Systems and Computing, vol. 1228, pp. 521–541. Springer (2020). DOI 10.1007/978-3-030-52249-0_35. URL https: //doi.org/10.1007/978-3-030-52249-0_35
2. Almeida, T., Hidalgo, J.M.G., Silva, T.P.: Towards sms spam filtering: Results under a new dataset. International Journal of Information Security Science 2(1), 1–18 (2013)
3. Anagnostou, A., Mollas, I., Tsoumakas, G.: Hatebusters: A web application for actively reporting youtube hate speech. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pp. 5796–5798. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden (2018). DOI 10.24963/ijcai.2018/841. URL https://doi.org/10.24963/ijcai.2018/841
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, May 7-9, 2015, Conference Track Proceedings. San Diego, California, USA (2015). URL http://arxiv.org/abs/1409.0473
5. Benites, F., Sapozhnikova, E.: Haram: A hierarchical aram neural network for large-scale text classification. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 847–854. IEEE Computer Society, USA (2015). DOI 10.1109/ICDMW.2015.14
6. Chen, J., Mao, J., Liu, Y., Zhang, M., Ma, S.: Tiangong-st: A new dataset with large-scale refined realworld web search sessions. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, November 3-7, 2019, pp. 2485–2488. ACM, Beijing, China (2019). DOI 10.1145/3357384.3358158. URL https://doi.org/10.1145/3357384.3358158Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17, pp. 512–515. AAAI Press, Montreal, Canada (2017)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Dinakar, K., Picard, R.W., Lieberman, H.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying (extended abstract). In: Q. Yang, M.J. Wooldridge (eds.) Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, pp. 4168–4172. AAAI Press (2015). URL http://ijcai.org/Abstract/15/589
9. Dramé, K., Mougin, F., Diallo, G.: Large scale biomedical texts classification: a knn and an esa-based approaches. J. Biomedical Semantics 7, 40 (2016). DOI 10.1186/s13326-016-0073-1. URL https://doi.org/10.1186/ s13326-016-0073-1
10. Fersini, E., Rosso, P., Anzovino, M.: Overview of the task on automatic misogyny identification at ibereval 2018. In: IberEval@ SEPLN, pp. 214–228 (2018)
11. Friedman, J.: Stochastic gradient boosting. department of statistics. Tech. rep., Stanford University, Technical Report, San Francisco, CA (1999)
12. Furini, M., Montangero, M.: Sentiment analysis and twitter: a game proposal. Pers. Ubiquitous Comput. 22(4), 771–785 (2018). DOI 10.1007/s00779-018-1142-5. URL https://doi.org/10.1007/s00779-018-1142-5
13. [14] Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: Z. Waseem, W.H.K. Chung, D. Hovy, J.R. Tetreault (eds.) Proceedings of the First Workshop on Abusive Language Online, ALW@ACL 2017, Vancouver, BC, Canada, August 4, 2017, pp. 85–90. Association for Computational Linguistics (2017). DOI 10.18653/v1/w17-3013. URL https://doi.org/10.18653/v1/w17-3013 Gao, L., Huang, R.: Detecting online hate speech using context aware models. In: RANLP (2017)
14. Geisser, S.: Predictive inference, vol. 55. CRC press (1993)
15. de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate speech dataset from a white supremacy forum. Proceedings of the 2nd Workshop on Abusive Language Online (ALW2) (2018). DOI 10.18653/v1/w18-5102. URL http://dx.doi.org/10.18653/v1/w18-5102

16. Haagsma, H., Bos, J., Nissim, M.: MAGPIE: A large corpus of potentially idiomatic expressions. In: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds.) Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, pp. 279–287. European Language Resources Association (2020). URL https://www.aclweb.org/anthology/2020.lrec-1.35/
17. Hoang, T., Vo, K.D., Nejdl, W.: W2E: A worldwide-event benchmark dataset for topic detection and tracking. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, pp. 1847–1850. ACM (2018). DOI 10.1145/3269206.3269309. URL https://doi.org/10.1145/3269206.3269309
18. Ibrohim, M.O., Budi, I.: Multi-label hate speech and abusive language detection in Indonesian twitter. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 46–57. Association for Computational Linguistics, Florence, Italy (2019). DOI 10.18653/v1/W19-3506. URL https://www.aclweb.org/anthology/ W19-3506Inc., M.: Kappa statistics for attribute agreement analysis. Available at https://support.minitab.com/ en-us/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/how-to/ attribute-agreement-analysis/attribute-agreement-analysis/interpret-the-results/all-statistics-and-graphs/ kappa-statistics/ (2021/04/17)
19. [22] Jirotka, M., Stahl, B.C.: The need for responsible technology. Journal of Responsible Technology 1, 100,002 (2020). DOI https://doi.org/10.1016/j.jrt.2020.100002. URL http://www.sciencedirect.com/science/article/pii/ S2666659620300020
20. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models (2016)
21. Karlos, S., Kanas, V.G., Aridas, C.K., Fazakis, N., Kotsiantis, S.: Combining active learning with self-train algorithm for classification of multimodal problems. In: IISA 2019, Patras, Greece, July 15-17, 2019, pp. 1–8 (2019). DOI 10.1109/IISA.2019.8900724. URL https://doi.org/10.1109/IISA.2019.8900724
22. Kim, S., Kim, D., Cho, M., Kwak, S.: Proxy anchor loss for deep metric learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 3235–3244. IEEE (2020). DOI 10.1109/CVPR42600.2020.00330. URL https://doi.org/10.1109/CVPR42600.2020.00330 [26] Krempl, G., Kottke, D., Lemaire, V.: Optimised probabilistic active learning (OPAL) - for fast, non-myopic, cost-sensitive active classification. Mach. Learn. 100(2-3), 449–476 (2015). DOI 10.1007/s10994-015-5504-1. URL https://doi.org/10.1007/s10994-015-5504-1 [27] Kumar, P., Gupta, A.: Active learning query strategies for classification, regression, and clustering: A survey. J. Comput. Sci. Technol. 35(4), 913–945 (2020). DOI 10.1007/s11390-020-9487-4. URL https://doi.org/10.1007/ s11390-020-9487-4
23. Ljubešic, N., Erjavec, T., Fišer, D.: Datasets of Slovene and Croatian moderated news comments. In: Proceedings ´ of the 2nd Workshop on Abusive Language Online (ALW2), pp. 124–131. Association for Computational Linguistics, Brussels, Belgium (2018). DOI 10.18653/v1/W18-5116. URL https://www.aclweb.org/anthology/ W18-5116