VADER meets BERT: sentiment analysis for early detection of signs of self-harm through social mining

Lucas Barros, Alina Trifan and José Luís Oliveira

DETI/IEETA, University of Aveiro, Portugal

Abstract

This paper describes the participation of the Bioinformatics group of the Institute of Electronics and Computer Engineering of University of Aveiro (BioInfo@UAVR) in the 2nd shared task of CLEF eRisk 2021. eRisk is an "Early Risk Prediction on the Internet" challenge whose tasks consist in analysing social media data and foster research on early detection of mental disorders. This year eRisk had 3 tasks, each focusing on a different disorder. This paper focuses on addressing the 2nd task, whose main objective is the early detection of users at risk of self-harming, based on their Reddit history. We addressed this issue by developing four supervised machine learning models that can classify such users. Our approaches are based on keyword extraction, sentiment analysis and word embeddings.

Keywords

social mining, early risk detection, mental health, self-harm,

1. Introduction

Suicides still remain one of the leading causes of death in many countries. In the USA, for instance, one can account for an average number of more than 40,000 suicides per year [1]. To handle this situation, it is important to develop and to disseminate public health surveillance tools that effectively help finding people who might be in risk of committing suicide. A possible driver for this path could be the exploration of data and textual information captured through clinical appointments. New algorithms and classification models can help researchers and clinicians identify in such datasets depression symptoms in early stages. However, this is not a simple task, due to the diversity of expressions, sentiments, absence of opinions, and many others [2].

Nevertheless, there are more data aside from clinical appointments. There is also a huge amount of user data online, such as: blogs, forum discussions, online reviews of products and services, queries to search engines and social media information. Taking this into consideration, the following question arises: how can large amounts of social media data inform Public Health? The goal of social media mining is to complement the clinic traditional information, not to replace it. This data can be used to generate hypothesis that connect variables to the state of mind of a user. Researchers argue that social media information will have its importance increased in public health in the future. Compared to traditional public health monitoring, social media-based monitoring is fast, cheap, and covers a large population. Social Media is also

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
lucasfrb45@ua.pt (L. Barros); alina.trifan@ua.pt (A. Trifan); jlo@ua.pt (J. L. Oliveira)

very helpful because it provides timestamps of users posts in social Media, which enables a chronological study about a user [3].

eRisk [4] is the "Early Risk Prediction on the Internet" challenge whose tasks are related to analysing social media data and it helps research on this topic. It has occurred in the last years with the participation of dozens of teams. Reading and analyzing those teams approaches to solve the tasks of the challenge will provide valuable learning on this topic.

This paper describes the participation of the BioInfo@UAVR team in the CLEF eRisk 2021 Task 2. The methodology and associate results are presented in this paper, as well as a discussion on future work. The rest of this paper is organized as follows: Section 2 overviews the current background in social data mining. The next two sections are dedicated to the description of the methodology and results obtained. The conclusion and discussion of possible improvements and future work are presented in Section 6.

2. Background

eRisk, as the name suggests, is a research initiative that promotes the scientific advance in this area [5]. It primarily focuses on the detection of mental health problems by using social data retrieved from social networks that can push forward new discoveries and insights. We are in need of new discoveries in this particular area. Over the years the results obtained in these tasks have proven that the research question to be addressed is a complex one with enough room for scientific improvement. What this tells us is that there is a lot of information that we do not know concerning the expression of these mental diseases and that there are still a lot of discoveries to be made on this subject.

As Nirmita et al. [6] point out, it becomes even more important to produce research on this particular topic given that depression rates have been increasing, specially with the COVID-19 Pandemic. The next graphic shows the results of a study made with USA adults about anxiety and depression before COVID-19 and after.

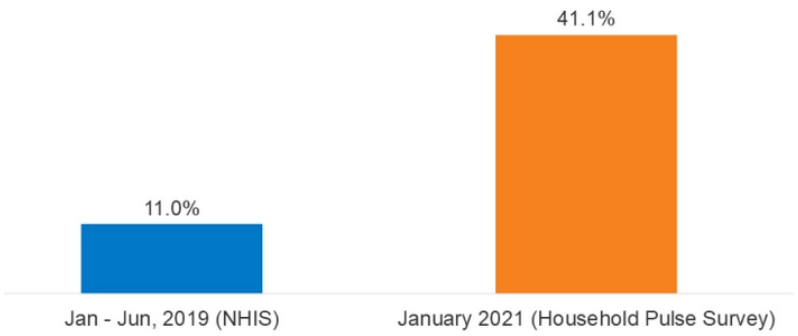


Figure 1: Pre-COVID vs COVID rates on anxiety and depression [6].

We decided to build on the steps from the last year's best performing team ILAB. ILAB used a BERT model to provide features to a machine learning classifier to predict the probability

of each subject being in a low mental health state. Their results wore the best in terms of F1 score, 0.67 and $F_{latency}$ 0.605. We decided to combine BERT features with VADER, a sentiment analyser. Due to some technical difficulties that we experienced in the testing phase and we were only able to analyse 91 out of the 2000 posts from each user which damaged our results.

3. eRisk - Task 2

Task 2 of eRisk consisted in processing data from users in the correct order, and detect signs of self-harm as soon as possible. There are 2 types of users in the dataset: users that at some point have harmed themselves (self-harm users) and users who have not (control users).

In this task, there are 2 phases: the training and testing phases. The training phase is about each team feeding the algorithms developed with data to optimize the parameters. In this phase, all the data is given at once. The data needs to be labeled, which means that the category of the user needs to be known, for the algorithm to learn how to distinguish between both categories. The testing phase is about each team, upon receiving a post from a specific user, making a binary decision regarding if the user is at risk of harming himself or not. If the decision is positive, then an alert is emitted and the decision is considered final. This means that it is no longer necessary to analyze further information of that specific user. If it is a negative decision, then that user's posts will continue to be analyzed. If every post is assigned a negative label than the user is not at risk of harming himself. In this phase, the data is given iteratively.

3.1. Metrics

Aside from general metrics such as Precision, Recall and F1, Losada et al.[7] proposed the $ERDE$ measure. This is an error measure that also takes into consideration how many posts are necessary until the algorithm detects a true positive case. In this way, algorithms that are slower in detecting true positive cases will have a greater error. Another used measure $F_{latency}$ proposed by Sadeque et al. [8], which combines the F measure (efficiency) and the delay in emitting alerts.

3.2. Dataset

To train and test each model's performance we used the eRisk dataset provided. The dataset comes with the following organization:

Each file is a xml file that contains all writings of a user. The file contains the user id, which uniquely identifies a user, and a list of writings of the user. Each writing has a title, a date, a text and an info field. The title is optional, so a lot of posts do not have one. The date has both the date and time when the writing was posted. The text represents the post itself, and the info field is just information of the type of writing - in this case, all writings are reddit posts.

In the eRisk test phase, the test data is not received all at once like in the training phase. In the test phase, posts from every user are received iteratively one by one through an online server. The next round of posts will only be delivered after a decision for all posts in the current round has been sent to the server.

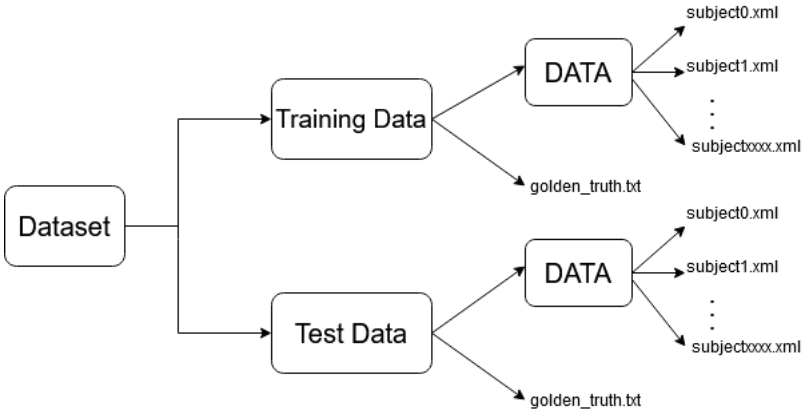


Figure 2: Conceptual organization of the dataset.

4. Methods

Our team worked on 4 different models for this task. Unfortunately, due to some technical issues during the training stage, only 2 were admitted due to a misunderstanding but they are all very similar. After analysing the approaches of several teams that participated in this task over the last years [9, 10, 11, 12, 13, 5] we decided to follow the approach of the team that achieved the best results. These were accomplished by the ILAB Team [13], with an approach based on BERT transformers. BERT [14] (Bidirectional Encoder Representations from Transformers) is a transformer based machine learning approach to solve Natural Language Processing (NLP) tasks. The BERT model is shown and described in Figure 3.

It should be noted that ILAB Team did not use the dataset that was provided by the eRisk challenge, but built their own, which may have had implications on the results obtained. In our approach we only used the data provided by the eRisk organizers.

BERT has 2 main components: a Tokenizer and a Model. The Tokenizer splits the string in tokens and the model receives the tokens and outputs the embeddings. These embeddings are vector representations of the sentence. Each token has a vector, so in the sentence *I like cats*, we will have a [5, 768] embeddings matrix. 5 because BERT adds 2 special tokens *CLS* and *SEP* for classification purposes. The embeddings are then used as input for a Support Vector Machine learning classifier that is previously trained with similar data and outputs a score for each of the 2 labels (0 for the control class and 1 for the self-harm class).

As classifier, our team chose a SVM (Support Vector Machine) since it is a very powerful algorithm to distinguish between classes. We also performed tests with the Adaboost classifier but no improvement was detected so we kept to the SVM.

4.1. First model

The first step is always a pre-process phase. In this case, the posts were lowercased and tokenized. After that, each post was fed to the BERT model described above in which the embeddings were obtained. As classifier, a SVM was used. There were also tests made with other classifiers, such

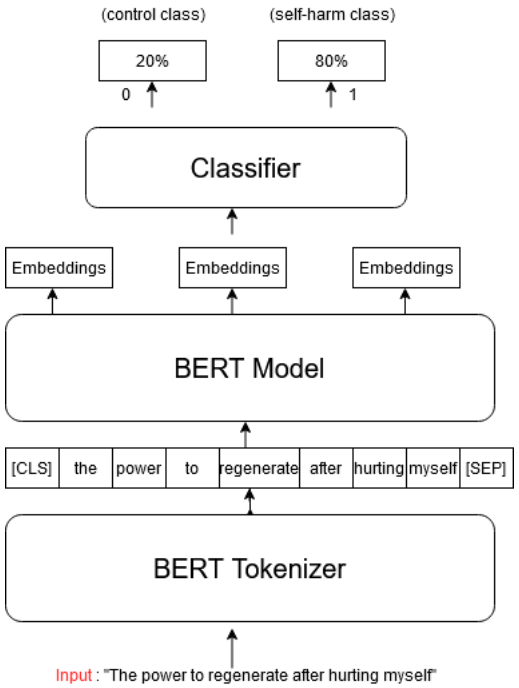


Figure 3: Illustration of the BERT architecture.

as Adaboost, but no improvements were obtained.

4.2. Second model

For the second model, a different kind of pre-processing was tested. Yake [15] extraction tool was used. Yake is a library to extract keywords from text. It is a very useful tool since, with the growth of information, it is not always viable to process huge amounts of text.

The results using Yake were not as expected though. A possible reason is because Yake extracts the important keywords. But what does important mean? The important keywords may very well depend on the problem. Yake extracts overall important keywords but this is a very specific problem. To solve this problem, like the NLP-UNED's Team showed, features like the number of first person pronouns can be used to distinguish between possible self-harm users and control users, and those are examples of words that algorithms like Yake will probably disregard. An extractor of keywords' algorithm could be very useful but it needs to be suited for the problem. Otherwise too much important information will be lost.

As an example, let us consider the post "I've been through some myself but I would never talk about it with other people, even if I was given a wide open opportunity to do so." taken from the eRisk test dataset.

This post was carefully chosen because it might indicate some mental health distress. However, the output from yake removes the important parts that might suggest that.

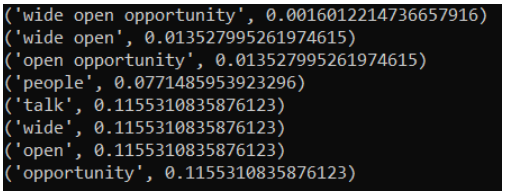


Figure 4: Yake output for the post "I've been through some myself but I would never talk about it with other people, even if I was given a wide open opportunity to do so."

4.3. Third model

This model was an updated version of the previous one, in which we introduced 2 new features: Emojis and sentiment analysis through VADER.

4.3.1. Emojis

First, we decided to consider the emojis from the posts. Emojis are a visual representation of an emotion, object or symbol[1]. Nowadays, every communication app, even the standard message app that comes with the smartphone, has an option that allows users to add emojis to the text. Since emojis are used to express emotion, that makes them very useful to determine the state of mind of a person. Two approaches were taken to add this feature to the model.

In the first approach, all the emojis in the emot library[2] were gathered. Then, for each post, an attribute emoji is created, which is a list of 0s where the number of 0s is equal to the number of different emojis. Then, the program finds emojis in the post and increments the corresponding 0 in the list. In the end, the lists of every post from the same user are summed and the resulted list is passed as a feature. This approach did not improve much the results obtained in the training stage, so a different one was pursued. This one works in a similar way but instead of creating lists, emojis are found in the text, they are replaced with the description of the emoji. As an example, the emoji *:-)* will be replaced by the corresponding description *joy*. If the original post is *Hi Andrew, nice to meet you :-)*, then the final post will be *Hi Andrew, nice to meet you joy.*

It may not make sense in terms of meaning, but the algorithm is not concerned with that. The word joy will make a lot of difference, not only because it will be associated with a happy feeling (control user), but also because the third feature used will be affected by these words, as explained next.

4.3.2. VADER

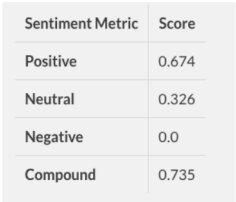
VADER [16] is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is very suited for this task. There is a Python module that makes it extremely easy to use.

It is important to point out that VADER is not 100% reliable. According to Pandey [**?**] there are a lot of tricky sentences. For example *The intent behind the movie was great, but it could*

[1] https://www.groovypost.com/howto/what-are-emojis-how-and-when-to-use-them/
[2] https://github.com/NeelShah18/emot

have been better. This sentence expresses both a positive and a negative sentiment. Similar, *The best I can say about the movie is that it was interesting*. In this example, interesting is a tricky word, it is not clear even for humans, if it means a positive or a negative sentiment. In real life situations, the tone of the conversation, or the facial expression of the person can help clarify it, but for an algorithm that analyzes text, those options are not available.

To give an example, let us consider the sentence *The phone is super cool.* By analysing this sentence, VADER gives the following output.



Sentiment Metric	Score
Positive	0.674
Neutral	0.326
Negative	0.0
Compound	0.735

Figure 5: VADER output for the sentence *The phone is super cool.*

The Positive, Negative and Neutral scores represent the proportion of text that falls into these categories. What this means is that 67% of the sentence gives a positive sentiment, 33% a neutral and 0% negative. Because it is a percentage, all of these summed up should be 1. The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1 (most extreme negative) and +1 (most extreme positive). The compound score turns out to be 0.75, that means a very high positive sentiment.

The emojis feature also affects VADER output. Let us consider the sentence from the example above, *Hi Andrew, nice to meet you :-)*. The following table shows the results given by VADER without the emojis characters, with the emojis characters and with the emojis translation.

Table 1
VADER output with the corresponding strings

Sentence	VADER compound
"Hi Andrew, nice to meet you"	0.4215
"Hi Andrew, nice to meet you :-)"	0.6249
"Hi Andrew, nice to meet you joy"	0.765

As it can be perceived, the emojis have a big impact on the sentence score given by VADER. The following diagram shows the various steps of the model presented.

4.4. Fourth model

The fourth model uses the same features as the third model. The only difference between the 2 is the usage of Yake in the pre-process phase. As it was concluded in the second model, the use of Yake does not improve the performance of the model, and the same was perceived in this one.

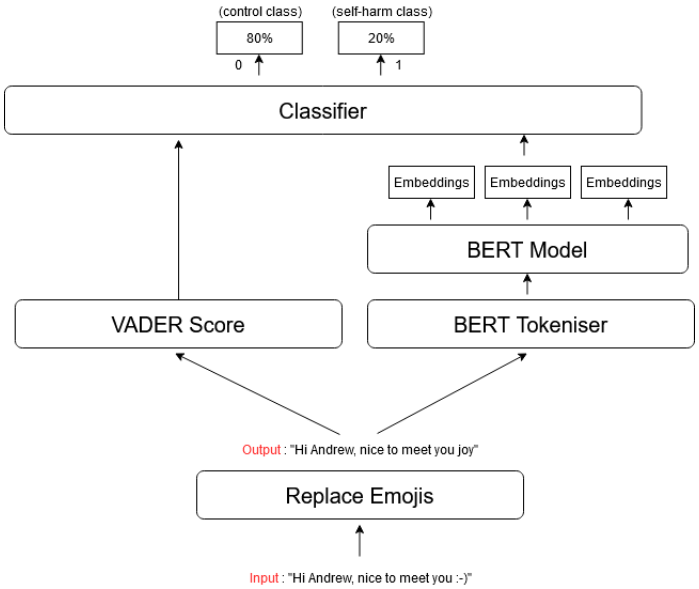


Figure 6: Pipeline of the third model proposed.

5. Results

Unfortunately, due to a technical mishap, the results of the last 2 models are not available. The results obtained in the first 2 runs are shown in the table bellow.

Table 2
Overview of the T2 official results.

Team Name	P	R	F1	ERDE5	ERDE50	latTP	speed	latency-weighted
Model 1	0.233	0.862	0.367	0.136	0.05	22	0.918	0.337
Model 2	0.274	0.789	0.407	0.128	0.047	22	0.918	0.374
Best Results	0.757	1.0	0.627	0.059	0.034	1	1.0	0.622

The best results row is the set of best measures in all submissions and not the submission that got the best results.

Before discussing the results, it is important to point out that only 91 out of around 2000 posts were processed, which is considerably low. The algorithm takes a long time to run and there were complications in the process.

As a result, the performance of our models was not optimal. The precision measure is significantly lower than the best achieved. Usually precision and recall have an inversely proportional behaviour so when one of them is really high, the other one is really low. The results achieved are an example of that, the recall measure is really high, but the precision is low, which means that our algorithm was able to find almost all of the self-harm users but a lot of the users classified as self-harm do not belong to this class.

Our latency results are also far from the best achieved, but we have to consider there might

be algorithms that only processed around 10 posts per user. In such cases, the decisions were emitted really fast. However, by processing so few posts and emitting decisions so fast, the other measures were all very low.

The team that got the best results overall was the UNSL team that got the best latency-weighted measure, after processing all 1999 posts. We believe our own results could have eventually improved if should we had processed more posts. In an offline testing environment we processed half of the official test corpus and our best performing model reached 0.447 in F1 latency-weighted.

6. Conclusions and future work

We presented in this paper the results of our team's participation in the eRisk2021 Task 2. As it was said, there were some complications in the process which damaged the results we obtained, which was unfortunate. The big amount of time the algorithm takes to process posts also influenced the results. There is a lot of potential in this work so, future work could be about improving the model, make it faster so that more posts can be processed and also explore other possibilities of features such as number of words, number of first pronouns for example. One important aspect to note regarding our approaches is that keyword extraction did not positively contribute to the results. However, the possibility of developing a keyword extractor specifically designed for this problem might be a possible approach for improving the results.

We recognize that social mining can have great use in selecting possible users that suffer from mental diseases, however given that best results are still not formidable, we have come to understand that only user's online data might be insufficient to get the better results possible since it gives very incomplete information about a person. The ideal scenario for this work would be a combination of both programmers and doctors to team up and develop a system in which in the first stage, the program selects the possible cases of mental illness automatically and in the second stage, with the clinical information and with a doctor's evaluation, make the final decision.

Acknowledgments

This work was supported by the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968 and by by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: DSAIPA/AI/0088/2020.

References

[1] P. Mortier, R. P. Auerbach, J. Alonso, W. G. Axinn, P. Cuijpers, D. D. Ebert, J. G. Green, I. Hwang, R. C. Kessler, H. Liu, et al., Suicidal thoughts and behaviors among college students and same-aged peers: results from the world health organization world mental health surveys, Social Psychiatry and Psychiatric Epidemiology 53 (2018) 279–288.

[2] M. J. Paul, M. Dredze, Social Monitoring for Public Health, 1, 2017.

[3] K. Loveys, P. Crutchley, E. Wyatt, G. Coppersmith, Small but mighty: Affective micropatterns for quantifying mental health from social media language, 2017, pp. 85–95. doi:10.18653/v1/W17-3110.

[4] M.-R. P.-L. D. E. . C. F. Parapar, J., Overview of eRisk 2021: Early Risk Prediction on the Internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association, CLEF 2021, Springer International Publishing, Bucharest, Romenia, 2021.

[5] A. Trifan, P. Salgado, J. L. Oliveira, Bioinfo@uavr at erisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_43.pdf.

[6] R. K. Nirmita Panchal, F. 2021, The implications of covid-19 for mental health and substance use, 2021. URL: https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/.

[7] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2016, pp. 28–39.

[8] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 495–503. URL: https://doi.org/10.1145/3159652.3159725. doi:10.1145/3159652.3159725.

[9] L. Achilles, M. Kisselew, J. Schäfer, R. Kölle, Using surface and semantic features for detecting early signs of self-harm in social media postings, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_42.pdf.

[10] M. Aragón, A. P. López-Monroy, M. M. y Gómez, Inaoe-cimat at erisk 2020: Detecting signs of self-harm using sub-emotions and words, in: CLEF, 2020.

[11] H. B. Hosseinabad, E. F. Ersi, A. Vahedian, Detection of early sign of self-harm on reddit using multi-level machine, in: CLEF, 2020.

[12] E. C. Ageitos, J. Martínez-Romo, L. Araujo, Nlp-uned at erisk 2020: Self-harm early risk detection with sentiment analysis and linguistic features, in: CLEF, 2020.

[13] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using bert-based transformers: ilab at clef erisk 2020, 2020. URL: https://early.irlab.org/, early Risk Prediction on the Internet : CLEF workshop, eRisk at CLEF ; Conference date: 22-09-2020 Through 25-09-2020.

[14] J. D. M.-W. C. Kenton Lee, Bert: Pre-training of deep bidirectional transformers forlanguage understanding, 2019. URL: https://arxiv.org/pdf/1810.04805.pdf.

[15] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, A. Jatowt, Yake! collection-

independent automatic keyword extractor, in: European Conference on Information Retrieval, Springer, 2018, pp. 806–810.

[16] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 8, 2014.