Using Surface and Semantic Features for Detecting Early Signs of Self-Harm in Social Media Postings

Linda Achilles, Max Kisselew, Johannes Schäfer, and Ralph Koelle

Institute for Information Science and Natural Language Processing,
University of Hildesheim, Hildesheim, Germany
{achilles,kisselew,johannes.schaefer,koelle}@uni-hildesheim.de

Abstract. This paper describes the University of Hildesheim submission to the CLEF eRisk 2020 shared task on detecting early signs of self-harm in social media posts. We introduce four systems that apply different methods trying to address this task and a fifth ensemble system that combines the four other systems. The first four systems make use of features of different types, such as time intervals between posts, the sentiment and semantics of the writings by using bag-of-words vectors and contextualized word embeddings in a neural network approach. The results show that while all our systems achieve a high recall, the focus of future work should be further improvement of the precision. All systems and the ensemble model achieve a comparable performance of $F_{latency}$ values in the range of 0.367 to 0.424.

Keywords: Self-harm detection · Risk detection · Natural language processing · Word vectors · Contextualized word embeddings · Deep learning · Ensemble system

1 Introduction

Since the advent of Social Media networks, they revolutionized the way people interact with each other and so attracted academic and industry researchers in different fields of expertise [3]. This way the content of those networks also became popular in the field of natural language processing (NLP). Text and meta information of posts have been used, for instance, to predict the Big Five personality traits of microblog users [1]. Later, different kinds of mental disorders came into focus of the research community. Facebook language was analyzed to predict a depression in a medical record [7], as well as postpartum depression [4] and Twitter data were evaluated to detect depressed users. The significance of this research led to the emergence of initiatives such as the CLEF eRisk group.

The Early Risk Prediction On The Internet (eRisk) lab[1], as part of the CLEF initiative[2], was first organized in 2017. The lab's main objective is to evaluate metrics, systems and data collections regarding Social Media users' safety and health on the internet. During the first workshop there was a pilot task on depression detection and a corresponding data collection was released [12]. During the following years more tasks, such as the early prediction of signs of self-harm or anorexia have been conducted.

In 2020 there were two tasks:

1. Early detection of signs of self-harm
2. Measuring the severity of the signs of depression

The data for the first task is released in rounds. In each round participating teams receive one social media post for each social media author in the collection. The task is to analyze these posts and return a binary decision [0, 1] for each author if he or she is at risk for self-harm in addition to estimating the level of self-harm represented by a score [0, 10]. A round ends when a team submits their results starting the next round.

In the second task the data is released as entire collection for each social media author at once. After processing these writings, the team's system has to fill in a standard depression questionnaire, the Beck Depression Inventory (BDI)[2], for each social media author.

This paper describes the approach of the *hildesheim* team in the first task of CLEF eRisk 2020, the early detection of signs of self-harm in social media posts. Section 2 and 3 give more insight into the data collection and the necessary pre-processing as well as an overview of the evaluation metrics in eRisk. Furthermore, we deployed five different systems, two of them making use of *surface-based* features (see Sections 4.1 and 4.2), two systems making use of *semantic* features (see Sections 4.3 and 4.4) and the fifth system being an ensemble system (see Section 4.5) combining the decisions and scores from the other four systems. Each system is evaluated separately in the shared task as runs 0 to 4 of our team. Since the shared task limited the maximum number of allowed runs per team to 5, we were not able to test different configurations of these systems. In Section 4 we present each system in more detail.

2 Data and Pre-processing

In 2019 the shared tasks' chunk-based data processing approach was changed to an item-by-item option to analyse the data. In 2017 and 2018 participating teams received a *chunk* of data, a sub-collection of the writings limited by a specific time period, for each social media author. From 2019 on they would receive the data post by post. This second way of processing the social media data would give the participating teams the possibility to make a decision about the state

[1] Website of the eRisk lab: https://erisk.irlab.org/

[2] Website of the CLEF initiative: http://www.clef-initiative.eu/

of a specific social media author at any point in time and it was continued in 2020 [14].

The data collection process is described in [12], the training data set in [13] and the test data set is described in [15]. For 2020 the training data set is a collection of XML files, one for each subject (social media author). Each individual XML file comprises a collection of an author's writings in chronological order. Each writing consists of *title*, *text*, *info* and *date* elements. The *title* and *text* elements represent the title and the content of each social media post, the *date* element the timestamp. The *info* tag gives information on the social network the data was retrieved from: The social network Reddit[3]. Both *title* and *text* elements can be empty if the other element contains text. Sometimes *title* as well as *text* are both filled with content.

The training data provided XML files for 340 subjects, 41 of which belonging to the self-harm group, 299 to the control group. The total number of writings in the self-harm group is 6,927 posts in contrast to 163,506 in the control group. The difference between the groups is also very significant in the average number of writings per subject: 169.0 in the self-harm group and 546.8 in the control group. According to the shared task organizers the average number of words per writing is in the self-harm group a slightly higher (24.8) than in the control group (18.8). The self-harm test data of the 2019 was reused for the 2020 challenge as training data [14]. Table 1 summarizes the details about the training data.

Table 1. Summary of the training data set for eRisk 2020 task 1 (Early detection of signs of self-harm).

	Self-harm	Control
Number of subjects	41	299
Number of submissions	6,927	163,506
Average number of submissions per subject	169.0	546.8
Average number of words per submission	24.8	18.8

During manual inspection of the training data, a few problems were identified. For example, some posts contained many URLs or Hashtags while others only consisted of HTML codes. The latter were partially identified as posts written in Cyrillic script, thus probably in a language different from English. These inconsistencies, typical for social media, required some pre-processing before the training data could be processed with the systems.

In a first step the posts were tokenized[4] splitting them into sentences and tokens. At the same time a post ID was generated from the subject ID and

[3] Website of the social media network Reddit: https://www.reddit.com/

[4] Using the standard NLTK tokenizer: https://www.nltk.org/api/nltk.tokenize.html

timestamp to keep the post structure also on sentence level. In a second step language-detection[5] was applied to the sentences.

We removed sentences in cases where the language detection algorithm raised an error. This way URLs were sorted out of the data. Additionally, hashtags and posts consisting only of HTML escape characters were also removed from the training data set. The remaining sentences were used for training of the semantic features. Some sentences contained urban slang words (e.g. *dawg*) and were consequently misclassified during language detection. Therefore, we decided to include all the posts where the language detection did not fail, even though not all sentences were classified to be English.

An additional pre-processing step of the training data was necessary for the time analysis system (see Section 4.1). The posts were sorted by subject and timestamp and the time difference between two posts was calculated. For both groups (self-harm and control group) mean and standard deviation were calculated. For the self-harm group the mean was approximately *71 hours (2 days 23 hours)*, meaning that each subject approximately posted every three days on average. The standard deviation was approximately *563 hours (23 days 11 hours)*. For the control group the mean was approximately *35 hours (1 day 11 hours)* and the standard deviation was approximately *333 hours (13 days 22 hours)*.

3 Metrics

Precision, recall and the F-measure, calculated referring to the self-harm positive group of social media authors [14], were used as evaluation metrics for eRisk since its advent in 2017. Besides these standard Information Retrieval metrics the organizers of eRisk also introduced a new metric, *ERDE* (Early Risk Detection Error), that involves the time that a system took to make a decision, based on the number of writings that were needed to come to this decision. *ERDE* allows to penalize late decisions made by the system, meaning that a *late* decision needs to process more writings than an *early* alert. The metric's range is [0, 1] meaning a low value corresponds to a good result of the system.

However, ERDE comes with several limitations [13], so that F_{latency}, as proposed by Sadeque and colleagues [16], was added to the set of evaluation metrics for eRisk. An optimal system's value is 1.

In addition to the decision-based evaluation metrics described above, the organizers introduced also ranking-based evaluation methodologies. The standard Information Retrieval measures *P@10*, *NDCG@10* and *NDCG@100* are also applied in eRisk.

[5] Python language-detection library langdetect: https://pypi.org/project/langdetect/

4 Systems

This section describes the four individual systems and the fifth ensemble system which we developed for the submissions for the eRisk 2020 shared task. Each system (1 - 5) was submitted as in separate runs (0 - 4).

Our systems had a run time of 72d 20h describing the time passed from the first to the last response to the server providing new Social Media posts. Technical problems within our systems led to this high number, however, no manual offline processing was involved. The systems work entirely automatically.

The systems 1, 2 and 4 were originally implemented returning a *decision = 1* when there were no more posts of an author in the current round, interpreting the last score and decision as *final* for this user. After the test stage of the shared task this understanding of the evaluation metrics turned out to not fully comply with the actual application of these measures of the share task.

4.1 System 1 - Time Analysis

One of our hypotheses is that social media users at risk for self-harm post less regularly than those in the control group. For the first run (run 0) a rather simple time analysis algorithm was deployed. The first step was to retrieve the timestamp from the round before the current round to calculate the time that passed between the two posts for each subject. For the calculation of the score that represents the estimation of the level of self-harm the standard deviation that was calculated during the training phase was used. The range from <0 days to 13 days 21:44:23 was mapped to the score range of 0 to 5 (lower half of being *at risk*). The second half was mapped from the first half's upper border to 23 days 10:51:57. Everything above this value was set to the highest possible score (10).

The mapping was done with the following functions for the groups:

$$f_{cg}(x) = \min(x \cdot \frac{5}{b_{cg}}, 5) \tag{1}$$

$$f_{sh}(x) = \min(x \cdot \frac{10}{b_{sh}}, 10) \tag{2}$$

While b_{cg} refers to the control group's upper time border (13 days 21:44:23), b_{sh} contains the self-harm group's value (23 days 10:51:57), both being converted to seconds. The variable x refers to the time difference in seconds between two posts. Thereby, f expresses the score of the level of self-harm risk in the range of 0 to 10.

For the calculation of the decision, the current round's score f was compared to the mean score of the previous rounds. If the difference between the current score and the mean score is greater than ±3 the system returns *decision = 1*.

There were cases were the algorithm caused an error: During the first round, since we need at least two writings for calculating a time difference between two posts, and at least one more case, where the timestamp of two successive posts of the same author were identical. To prevent the system from causing an error we caught these cases by setting the default time difference to ten seconds.

4.2 System 2 - Sentiment Analysis

For the second run (run 1) we developed a method based on sentiment analysis. We used *VADER* (Valence Aware Dictionary and sEntiment Reasoner) [10], a lexicon-based sentiment algorithm due to its ability to process elements that are special for Social Media posts, such as emoticons as an expression of mood or words written in capital letters as a sign of emphasis.

VADER comes with a *compound* score which is described by the authors as a normalized and weighted composite score[6]. The compound score ranges between -1 (extreme negative sentiment) and +1 (extreme positive sentiment). Posts with sentiment values in the range from -0.05 to +0.05 are considered to be of neutral sentiment. The sentiment algorithm was applied on post level on the pre-processed training data. A histogram of the distribution of posts with specific sentiment values is shown in Figure 1. Both groups (self-harm and control group) are visualized separately. At first glance both groups have a similar distribution of sentiment compound values such as for instance the most posts are located in the neutral segment. However, the shapes of the histograms differ at the margins. The control group appears to have a decreasing number of extreme positive and extreme negative posts compared to the self-harm group. Therefore, we hypothesize a correlation between extreme sentiment values and the author's susceptibility to self-harm.

For the test stage, the sentiment method therefore functions as follows: VADER is applied on post level to calculate the compound sentiment value. In a next step, the compound sentiment value is mapped to the score for the estimation of the level of self-harm by multiplying the compound value by 10. A negative compound value is first transformed into a positive number by multiplying it by -1.

For the binary decision a mean score of the previous rounds is calculated. The algorithm then compares the mean score of the previous rounds with the score of the current round. If the difference between the current score and the mean score is greater than ±3 the system returns *decision = 1*, because we assume the author to be prone to self-harm, becoming apparent through a less stable sentiment.

4.3 System 3 - Distributional Semantics

For our third system (run 2) we implemented a system based on distributional semantics to decide as early as possible whether a post indicates signs of self-harm. To this end, our method computes the semantic similarity between vector representations of previously unseen Reddit posts and an abstract vector representation exhibiting the semantics of a typical post showing signs self-harm

[6] VADER on Pypi.org: https://pypi.org/project/vader-sentiment/

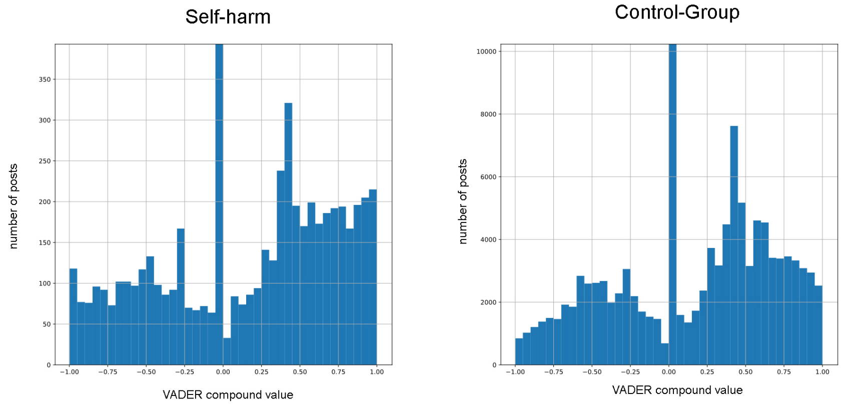

Fig. 1. Distribution of posts in conjunction with VADER sentiment compound value.

behavior. If the semantic similarity exceeds a predefined threshold, the system raises an alarm for the subject who wrote the corresponding post.

Our system employs an approach based on the framework of distributional semantics [19]. This framework is based on the distributional hypothesis which states that words occurring in similar contexts tend to have similar meanings [9,8,5]. This hypothesis is operationalized by representing the meaning of target words as multidimensional vectors, also referred to as *word vectors* or *semantic vectors*. In this work we use the bag-of-words model where each vector dimension stands for a particular context word and the numerical value in that dimension signifies how often that context word co-occurs within a window of n words around the target word in a specific text corpus.

Data. As described in Section 2 the training data contains a history of posts for 41 subjects for whom self-harm alarm has been raised and 299 healthy subjects. To develop our model on data balanced in terms of the distribution of self-harm and healthy subjects' histories of writings, we randomly selected 41 subjects from the set of healthy subjects. Together with the 41 subjects prone to self-harm this results in a total of 82 histories of posts.

Vector representation of Reddit posts. Our word vectors are built on a concatenation of the lemmatized British National Corpus (BNC)[7] and ukWaC corpora[8]. These corpora are used in many related works to build vector space models as they contain a wide range of text genres from different domains. The corpus contains in total around 2.36 billion tokens. We create 10000-dimensional bag-of-words vectors for the 28,000 most frequent content words in the corpus and set the window size for counting co-occurring context words around each

[7] British National Corpus (BNC): http://www.natcorp.ox.ac.uk

[8] UKWaC corpora: http://wacky.sslmit.unibo.it

target word to 5 words to either side of the target word. The vector dimensions are the 10,000 most frequent content words from the corpus. For each Reddit post that is used for the development of our system and during the test stage of the shared task, we compute an average semantic vector as follows: Assuming that W_p is the set of all words appearing in post p, we compute the vector $\overrightarrow{p}$ representing the meaning of post p as the centroid of all word vectors $\overrightarrow{w}$ for words w from W_p:

$$\overrightarrow{p} = \frac{1}{|W_p|} \sum_{w \in W_p} \overrightarrow{w} \tag{3}$$

Vector representation of susceptibility to self-harm. A manual inspection of writings from the 41 subjects who are prone to self-harm according to the golden truth revealed that in many cases the last post written by these subjects is particularly characteristic for the expression of self-harm behavior. Thus, we compute a special *self-harm vector* $\overrightarrow{s}$ indicating actions or thoughts of self-harm by averaging post vectors representing the last post from each of the 41 subjects prone to self-harm.

To tune our system, we computed the cosine similarities between the first five posts of each subject in our balanced data set consisting of 82 histories of posts and the *self-harm vector* $\overrightarrow{s}$. We found 0.985 to be a good threshold to distinguish posts from persons prone to self-harm and those who are not. Therefore we use this threshold for the test stage in the CLEF eRisk 2020 shared task.

During the test stage we compute a post vector for each incoming post as we did during the training phase. Then we compute the cosine similarity between each of these post vectors and the pre-computed *self-harm vector* $\overrightarrow{s}$. If the cosine similarity exceeds the threshold of 0.985, an alarm for that post is raised and the label 1 is assigned to the subject who wrote that post. The score expressing susceptibility to self-harm expressed in post p is computed using the following linear function:

$$\text{score}(p) = \cos(\overrightarrow{p}, \overrightarrow{s}) \cdot 10 \tag{4}$$

where $\overrightarrow{p}$ is the post vector and $\overrightarrow{s}$ the pre-computed *self-harm vector*.

4.4 System 4 - Neural Network

Automatically learning to estimate a person's risk of self-harm solely based on their social media postings is a complex task. Therefore, an automatic system has to be able to consider a variety of textual features. Manually defining a feature set which incorporates a comprehensive amount of indicators transpired to be a tedious task (cf. Systems 1-3 utilizing diverse feature categories). Thus, we decided to use a neural network to automatically learn such features. In this section we describe the structure of our deep learning model, how we train it on the available data set and how we predict scores for new instances. The overall architecture of our neural network system is shown in Figure 2.

In the context of the shared task, we formulate the classification task as follows: an instance (input data point) consists of a sequence of posts of a single

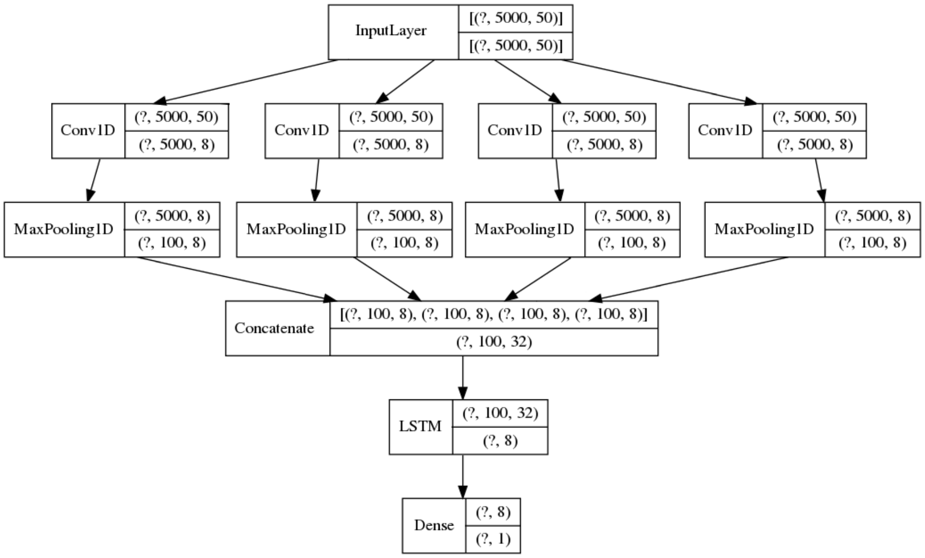

Fig. 2. Architecture of our neural network system. Each box corresponds to a neural network layer while the values on the right side of each box display the dimensions of the input (values at the top) and output (values at the bottom) tensor of each layer ('?' is a placeholder for the variable batch size).

author (where each post is represented as a string) and should be classified as a binary decision, expressing if the author tends toward self-harm (yes/no).

Model architecture. As numerical representation of the posts' text, we used *BERT* [6] embeddings which has in recent years proven to be an efficient method of represent meaning in contextualized word embeddings. We used the pre-trained embedding model *BERT-Base* ('BERT_uncased_L-12_H-768_A-12')[9] [18]. As dimensionality of the embeddings we selected 50 features. To keep the size of our instances computationally feasible, we limited the post history to the 100 most recent postings and the length of each post to the first 50 words (using pre-padding in both cases when fewer postings/words are available). Thus, an instance consists of 5,000 *BERT* word embeddings of size 50. These values are also shown in the input layer of the network in Figure 2.

Subsequently, we apply a convolutional neural network (CNN) to automatically select features which can be indicative of self-harm from the post history. Our CNN sub-network consists of four parallel branches, each being a sequence of a convolutional layer, a max pooling layer and a dropout layer. CNNs are efficient feature selectors which can be applied to learn complex features based on semantic representations, cf. [17]. We decided to use four different convolutions (see the branches in Figure 2) to detect 8 features (by using 8 filters in the layer) from word n-grams of different sizes n from 1 to 4. Thus, the first convolution learns unigram features from the embedding of the post history of the subject,

[9] The used BERT model is available on https://github.com/google-research/bert

the second convolution learns bigram features, etc. In each branch, the subsequent max pooling layer is applied to retain only the maximum feature values for each post, resulting in a 100x8 matrix (8 features for each of the 100 most recent postings) in each branch. A dropout layer (with a probability of 0.25) then introduces regularization to the model during training which improves its ability to generalize to new data instances. Finally, the output of the branches is simply concatenated resulting in 32 features for each of the 100 most recent postings as shown in the layer named *Concatenate* in Figure 2.

To compute an assessment of the overall self-harm risk of a subject based on these high-level features of their most recent history, we process the output of the *Concatenate* layer using a recurrent neural network (RNN) with long-short term memory (LSTM) cells (see *LSTM* layer in Figure 2). This neural network component reads the features of these posts as a sequence while updating an internal status (which here should model the risk of self-harm) after each step (post). Finally after processing the entire sequence, the RNN returns an overall assessment for the given post history; here, we set the dimensionality of the output space to 8.

As final layer we use a densely connected layer with a sigmoid activation function to predict an output between 0 and 1 which can be interpreted as the probability of the subject being prone to self-harm. With this configuration, the model has a total of 5,353 trainable parameters.

Training. We used the data available in the *eRisk 2019 Text Research Collection* [12] to construct training instances as follows. As our model expects a chronological list of posts per subject as input and the data set consists of the entire post history of only 340 subjects, we decided to use multiple random samples of subsets of this data. We constrain the samples to be between 10 and 100 posts in length while we consider both titles and text comments as individual posts. With this method we extract 3,705 samples from the data set and select 90% of those as our training data and the remaining 10% as validation data.

To cope with the class imbalance (in the original data set, only 41 of 340 subjects are prone to self-harm), we use class weights during training based on their inverse frequency in out training set. Thus, errors on classifying positive instances are valued approximately 21 times higher than errors on classifying negative instances. Additionally, to avoid over-fitting, we apply early stopping using the validation set which stops training when the monitored binary cross-entropy score for the validation set does not improve for several epochs. Finally, we optimize the network with a training batch size of 64 and a binary cross-entropy loss using the Adam optimizer [11] in 12 epochs over the training data.

Prediction. Our neural network model returns a probabilistic prediction $p(s)$ for each unseen instance expressing the likelihood of the subject s being prone to self-harm. To conform to the format of the shared task, we reformat this score for each subject into a decision using Equation 5 and a self-harm score $sh(s) = 10 \cdot p(s)$.

$$decision(s) = \begin{cases} 0 & \text{if } 0.1 < p(s) < 0.8 \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

Using this method, we set the decision value to 1 only in cases when the model is extremely confident for a negative prediction (values below 0.1)[10] or highly confident for a positive prediction (values above 0.8)[11].

4.5 System 5 - Ensemble System

Our ensemble system uses the individual predictions of the above-mentioned four systems (System 1-4) in each run to compute a single combined decision. For each subject this system predicts a decision value of 1 if at least two of the individual systems assigned a 1. If only one or none of the systems predicted 1, the ensemble system returns 0. The self-harm risk score for each subject is always computed as an average of the scores of all four systems in each run.

5 Results

Table 2 shows the overall decision-based results of team *hildesheim* in all five runs on the eRisk 2020 test data. The run ID corresponds to our systems as described in Section 4.

Our systems rank high in terms of recall. Two systems achieve $R = 1$, but also the other three runs are highly ranked. Corresponding to these results the precision of all five runs is comparably low. Our best system achieves a precision of $P=$ 0.297 which shows that approximately every third decision is correct. In such a sensitive domain we consider this to be an acceptable result, but we aim for a higher precision in future eRisk participation of our team.

For ERDE_5 one of our systems (System 3) performs well with a low value of $\mathrm{ERDE}_5 =$ 0.237. After processing 50 posts all our systems improve, e.g. our best system achieves an ERDE_{50} result of 0.185.

Regarding the latency-weighed F measure $\mathrm{F}_{\mathrm{latency}}$ our systems show average performance. Our third system performs best with a value of $\mathrm{F}_{\mathrm{latency}} =$ 0.424, close to the results of our other proposed systems.

The ranking-based evaluation results are listed in Table 3. We did not consider the results after the first writing to be meaningful for our approaches, because some systems are based on a history of writings and especially for the time algorithm at least two writings are necessary to make a decision. For that reason, we think that the results for 100 writings and 500 writings are more interpretable. However, the ranking-based results remain rather constant for 100 and 500 writings with only slight changes.

[10] We chose this very strict threshold to only exclude subjects which have shown several clear signs of **not** being prone to self harm.

[11] Here we chose a weaker threshold (in comparison to the threshold for negative cases) to catch subjects which only have shown signs of being prone to self harm at some point, though those still being clear signs.

Table 2. Our decision-based results of task 1: Early detection of signs of self-harm, in comparison to the best results achieved in the shared task for each metric.

System	P	R	F1	$ERDE_5$	$ERDE_{50}$	$F_{latency}$
1	0.248	1.000	0.397	0.292	0.196	0.397
2	0.246	1.000	0.395	0.304	0.185	0.389
3	0.297	0.740	0.424	0.237	0.226	0.424
4	0.270	0.942	0.420	0.400	0.251	0.367
5	0.256	0.990	0.406	0.409	0.210	0.389
Best	0.913	1.000	0.754	0.134	0.071	0.658

Table 3. Our ranking-based results of task 1: Early detection of signs of self-harm, in comparison to the best results achieved in the shared task for each metric.

	100 writings			500 writings		
System	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
1	0.4	0.43	0.42	0.5	0.53	0.42
2	0.5	0.48	0.49	0.5	0.54	0.57
3	1.0	1.00	0.69	1.0	1.00	0.68
4	0.1	0.07	0.13	0.1	0.06	0.11
5	1.0	1.00	0.62	1.0	1.00	0.69
Best	1.0	1.00	0.83	1.0	1.00	0.84

The complete list of the evaluation results of all participating teams is published in [15].

6 Conclusion and Future Work

This paper describes the University of Hildesheim submission to the CLEF eRisk 2020 shared task on detecting early signs of self-harm in social media posts. We presented four systems that apply different methods trying to solve this task and a fifth ensemble system that combines the decisions of the four former systems. The first four systems analyze social media posts taking into account different types of features, such as time intervals between posts, sentiment values and semantic representations. While all our systems achieve a high recall, they struggle to yield a comparable precision.

We expected our ensemble system to perform better than the other systems since it raises an alarm only if several other systems do so as well. However, as the results show, the ensemble system is more likely to balance out the predictions of the other four systems which becomes apparent by being ranked among the other four systems. Therefore as future work we intend to implement a more sophisticated ensemble model that is able to incorporate the strengths of the other four systems, in particular when those are very confident in their decisions.

This could be accomplished by weighting the contributions of the single systems in the ensemble system. Investigating different settings for the ensemble system is a promising avenue towards a solution to the early detection of self-harm risk in social media postings.

References

1. Bai, S., Hao, B., Li, A., Yuan, S., Gao, R., Zhu, T.: Predicting Big Five Personality Traits of Microblog Users. In: V. Raghavan, X. Hu, C.L.J.T. (ed.) Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). vol. 1, pp. 501–508. IEEE Computer Society, Washington D.C., USA (November 2013)
2. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An Inventory for Measuring Depression. Archives of general psychiatry **4**(6), 561–571 (1961)
3. Boyd, D.M., Ellison, N.B.: Social Network Sites: Definition, History, and Scholarship. Journal of computer-mediated Communication **13**(1), 210–230 (2007)
4. De Choudhury, M., Counts, S., Horvitz, E.J., Hoff, A.: Characterizing and Predicting Postpartum Depression from Shared Facebook Data. In: S. Fussell, W. Lutters, M.R.M.M.R. (ed.) Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. pp. 626–638. CSCW 2014, Association for Computing Machinery, New York, NY, USA (February 2014)
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the Association for Information Science **41**(6), 391–407 (1990)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
7. Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D.A., Schwartz, H.A.: Facebook Language Predicts Depression in Medical Records. Proceedings of the National Academy of Sciences (PNAS) **115**(44), 11203–11208 (2018)
8. Firth, J.: A Synopsis of Linguistic Theory 1930-1955. Studies in Linguistic Analysis pp. 1–32 (1957)
9. Harris, Z.: Distributional Structure. Word **10**(23), 146–162 (1954)
10. Hutto, C.J., Gilbert, E.: Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In: Adar, E., Resnick, P., Choudhury, M.D., Hogan, B., Oh, A.H. (eds.) Proceedings of the 8th International AAAI Conference on Weblogs and Social Media. AAAI 2014, AAAI Press, Palo Alto, CA, USA (June 1-4 2014)
11. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Bengio, Y., LeCun, Y. (eds.) Proceedings of the 3rd International Conference on Learning Representationss. ICLR 2015 (May 7-9 2015), http://arxiv.org/abs/1412.6980
12. Losada, D., Crestani, F.: A Test Collection for Research on Depression and Language Use. In: K. Balog, L. Cappellato, N.F.C.M. (ed.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 7th International Conference of the CLEF Association. pp. 28–39. CLEF 2016, Springer Intrnational Publishing, Cham, Switzerland (September 5-8 2016)

13. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk at CLEF 2019 Early Risk Prediction on the Internet (extended overview). In: Cappellato, L., Ferro, N., Losada, D.E., Müller, H. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Proceedings of the 10th Conference and Labs of the Evaluation Forum. No. 2380 in CEUR Workshop Proceedings (September 2019), http://ceur-ws.org/Vol-2380/paper_248.pdf
14. Losada, D.E., Crestani, F., Parapar, J.: eRisk 2020: Self-harm and Depression Challenges. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. Proceedings of the 42nd European Conference on IR Research. pp. 557–563. ECIR 2020, Springer International Publishing, Cham, Switzerland (April 14-17 2020)
15. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2020: Early Risk Prediction on the Internet. In: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (ed.) ”Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association. CLEF 2020, Springer International Publishing, Cham, Switzerland (September 22-25 2020)
16. Sadeque, F., Xu, D., Bethard, S.: Measuring the Latency of Depression Detection in Social Media. In: Proceedings of the 11th ACM International Conference on Web Search and Data Mining. pp. 495–503. WSDM 2018, Association for Computing Machinery, New York, NY, USA (February 2018)
17. Schäfer, J., Burtenshaw, B.: Offence in Dialogues: A Corpus-Based Study. In: R. Mitkov, G.A. (ed.) Proceedings of the International Conference Recent Advances in Natural Language Processing: Natural Language Processing in a Deep Learning World. pp. 1085–1093. RANLP 2019, INCOMA Ltd., Varna, Bulgaria (September 2-4 2019). https://doi.org/10.26615/978-954-452-056-4_125
18. Turc, I., Chang, M.W., Lee, K., Toutanova, K.: Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. arXiv preprint arXiv:1908.08962v2 (August 2019)
19. Turney, P.D., Pantel, P.: From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research **37**(1), 141–188 (2010)