Temporal Word Embeddings for Early Detection of Signs of Depression*

Manuel Couto[1,*,†], Anxo Pérez[1,†] and Javier Parapar[1,†]

[1]IRLab, CITIC, Computer Science Department, University of A Coruña, A Coruña, 15071, Spain

Abstract

Depression is one of the most debilitating mental health diseases. Detecting the presence of depressive symptoms in the early stages of the disease is essential to reduce further consequences. As the study of language and behaviour is a pivotal component in mental research, social network content positions itself as a helpful tool. This paper introduces a general framework to analyze variations in the individual's use of language over time on social media. We present a novel approach using temporal word representations to quantify the magnitude of words movements. This framework allows us to evaluate if words evolution can reveal the presence of depressive tendencies. We adapted different temporal word embedding representations to our framework and assessed them in Reddit benchmark datasets. Our results achieve high competitiveness compared with state-of-the-art methods, showing the potential that time-aware word representation models can bring to early detection scenarios.

Keywords

Early Depression Detection, Social Media Analysis, Word Embeddings, Temporal Word Representations

1. Introduction

Depressive conditions affect around 300 million people [1], being a primary reason for suicide in the US [2] and disability cases worldwide [3]. The number of depressive cases has been increasing over the years, with the effects of the Covid-19 pandemic causing a notable escalation in recent months [4]. This trend is especially worrying given that the majority of patients are adolescents and young adults [5]. In addition, the disease's effects extend beyond the sanitary scope; generating economic, social, and labour impacts [6]. All these factors have raised the awareness of many institutions, which are starting to address this public health issue.

A growing body of evidence shows that proper interventions can significantly reduce the adverse effects of the disorder [7]. More specifically, several studies have demonstrated how early identification of these disorders is crucial to minimize their consequences [8, 9]. In this regard, the work of mental health researchers has focused on the relationship between mental disorder stages and language use [10, 11]. With the popularization of social media platforms, researchers showed the potential of these sources for analyzing the language that the users

CIRCLE (Joint Conference of The Information Retrieval Communities in Europe), July 04-07 2022, Toulouse, Fr

*Corresponding author.

†These authors contributed equally.

manuel.couto1@udc.es (M. Couto); anxo.pvila@udc.es (A. Pérez); javier.parapar@udc.es (J. Parapar)

https://www.dc.fi.udc.es/~parapar (J. Parapar)

0000-0002-0593-7199 (M. Couto); 0000-0002-0480-006X (A. Pérez); 0000-0002-5997-8252 (J. Parapar)

employ in that context [12]. People use social media to express their thoughts and feelings and to talk with peers that have shared interests [13]. Furthermore, for most people, the anonymity that these platforms provide may disinhibit them when communicating [14]. In many cases, this results in a more personal and emotional language, leading to self-disclosure of mental health status [15, 16, 17].

In this context, initiatives such as CLEF Early Risk (eRisk) emerged as an online competition focusing on early risk prediction on the internet. Five editions have been held since 2017 [18, 19, 20, 21, 22]. In 2017, the organizers introduced the pilot task on *Early Detection of Depression* (EDD) [18]. The task consisted of sequentially processing pieces of users' posts to detect early traces of depression. Organizers proposed the task in two editions (2017 and 2018, and the third edition, 2022, is currently ongoing). The challenge provides train and test data, composed of the history of writings of positive and negative Reddit users. The user's writings are sorted chronologically and split into ten subsets or chunks for each user. Each chunk contains 10% of the posts. This division into chunks is especially relevant, as participants will process each chunk sequentially to detect pieces of evidence as soon as possible. To properly consider the earliness of the decision of participant systems, organizers defined time-aware classification metrics.

This paper proposes a novel approach to predict and detect early depressive symptoms. Our framework tries to capture changes in individuals' language use over time. The idea is based on the premise that, when people develop a depressive disorder, they usually change their use of language. We intend to seek patterns in how individuals' language evolution can manifest changes in their mental health states. For example, in an individual with a healthy mental condition, the word *life* appears in the context of words that are related to positive or neutral terms (*great, going on, ok, good*). Conversely, if that same individual begins to develop depressive symptoms, *life* would start appearing in alarming contexts (*end, ruined, miserable*) in the progressive temporal periods.

In particular, we introduce the use of *temporal word embeddings for early detection of signs of depression.* Temporal models have been traditionally used to capture the language and cultural evolution, focusing on meaning shifts in long periods (decades to centuries). For example, the word *gay* originally meant cheerful or pleasant, and only after the decade of 1990 acquired the primary meaning of homosexuality [23]. Unlike all previous work on temporal word representations, we propose using these models as a resource to detect mental health shifts in the short term. We present a general framework for the *EDD* task. It calculates the magnitude of word movements over different time slices. These movements over time are the features considered during the training and inference classification phases. The model computes the movement of a word as the distance between the temporal word embedding for the user in a particular chunk and a reference static word embedding that captures general language.

We tested two temporal embeddings models. (1) *Temporal Word Embeddings with a Compass* (TWEC) [24], built on the word2vec architecture. This model generates one unique embedding per word, ignoring the local context where the word appears. (2) *Dynamic Contextualized Word Embeddings* (DCWE) [25], which is based on pre-trained language models (PLMs), such as BERT [26]. PLMs provides *contextualized* embeddings, i.e, capturing context-specific vectors. We evaluated our methods in the *EDD* eRisk benchmark collections of 2017 and 2018. Our proposal provides an interesting exploratory window of how individuals' usage of words evolves under

certain mental states. Despite not using any linguistic resource or hand-crafted features specific to depression, we obtained competitive results among the participants in both editions. Our findings suggest that temporal representation methods can be a reliable solution for capturing the evolution of depressive symptoms. In this regard, the adaptability of our framework also makes it easy to transfer these methods to other related disorders without additional effort.

The remainder of the paper is structured as follows. Section 2 provides a review of related work about depression detection on social media and the use of temporal models. Section 3 presents our proposed framework. First, we introduce the general idea and then the actual approaches implemented. We describe the objective task and the collections used in our research, along with the experimental configuration and results in Section 4. Finally, Section 5 discusses the main conclusions derived from this work and proposes future lines of work.

2. Related Work

Researchers from the psychology, medicine or linguistics fields have performed extensive work on topics related to mental health conditions and language. These works found interesting connections between language use and mental health [27, 28, 29]. In recent years, social media has become an essential source to provide data related to mental health disorders [30, 31, 32, 33, 34]. This amount of valuable information allows researchers to analyze user language and behaviour. Many studies have investigated how social network content can predict and identify mental conditions. As a result, computational methods have been applied to detect the presence of disorders like schizophrenia [35], suicidal ideation [36, 37] or eating disorders [38].

Data coming from social platforms like Twitter [39], Facebook [40] or Reddit [31] has been used to model the characteristics of mental disorders. De Choudhury et al. [41, 42, 43] made pioneer contributions to the field, extracting relevant features in the language of depressive people, such as the existence of high self-attentional focus and negative emotions [43]. Similarly, several researchers started to investigate a wide range of distinctive features: linguistic style, emotional expression, semantic, lexicon-based, social network properties, posting activity, among others. For instance, Ortega-Mendoza et al. [44] analyzed the relevance of the personal statements for depression detection in Reddit. Other studies [39, 45] also found that emotions play a relevant role in depressive writings. In this context, shared tasks emerged to promote the research on mental health topics based on social media. The two evaluation forums that have become standard benchmarks in this field are The Early Risk Prediction on the Internet (eRisk)[1] and the Computational Linguistics and Clinical Psychology (CLPsych)[2]. This section narrows the analysis to works on the former initiative regarding early risk detection.

2.1. eRisk Shared Task: Early Detection of Depression (EDD)

Detecting depression in the early stages has great importance in reducing the disorder's impact. Therefore, when building automatic alert systems, the sooner the alert emission on the temporal sequence, the better. For this reason, eRisk organized the first shared task on early risk detection

[1] https://erisk.irlab.org/
[2] https://clpsych.org/

of depression in 2017. To tackle the time aware nature of the task, the organizers proposed *Early Risk Detection Error* (ERDE)[3], a new metric which considers both accuracy and the delay of the predictions. In the first edition of the task, participant systems used traditional machine learning models such as bag-of-words, topic models, standard classifiers, or neural networks approaches [18]. They also experimented with feature extraction of different depression lexicons [46], sentiment and emotions, or semantic analysis [47], among others. More recently, Burdisso et al. [45] proposed a text classifier, called SS3. This new model incorporates three key aspects: incremental classification, support for early classification and explainability. SS3 maps words to categories, estimating a degree of confidence that a word exclusively belongs to a category. The authors evaluated their methods in the eRisk 2017 collection using positive and negative categories, obtaining remarkable results compared to the participants.

In the second edition, participants proposed new approaches leveraging word embeddings and Deep Neural Networks to capture the semantics of the individuals and combined them with the best performing features of 2017 [19]. Most of the participants improved their methods from 2017. Furthermore, some of them also developed classifiers supporting incremental classification. For instance, in Funez et al. [48], the solution remembers all the historical information until the present chunk. More recently, Ramiandrisoa and Mothe [49] presented an approach that combines two types of sentence representations: (1) based on previous features extracted from users' writings (some specifically designed for depression) and (2) based on vector embeddings. Combining both approaches (ModComb) improved the results from the top participants in some of the official metrics in 2018.

2.2. Temporal Word Embeddings (TWE)

Lately, researchers have adopted neural language models as the most common way to capture semantic features from texts. These models encode the word meanings as vectors, known as word embeddings. Words that appear in similar contexts will be closer in a multi-dimensional space. Traditional word embedding architectures, including Word2vec [50] and GloVe [51], are formulated as static models. Static models use a single *global* embedding for each word, ignoring the variability of word meanings. This fact implies two main restrictions. Static architectures do not capture how the meaning vary across: (i) linguistic contexts (i.e., to model polysemy) and (ii) extralinguistic contexts (i.e., to model temporal variations).

(i) The introduction of context-dependent approaches addressed the first limitation. Most contextualized embeddings come from modern transformers architectures, such as ELMo [52], BERT-based models [26], GPT [53] or T5 [54], resulting in performance improvements on a variety of NLP tasks.

(ii) A growing body of research focused on the second limitation, creating approaches known as Temporal Word Embedding Models (TWE) or Diachronic models. However, learning embeddings in temporal contexts is a complex task. Traditional temporal methods relied on the use of static neural models. Due to the stochastic nature of neural networks, the models trained in different temporal moments will have vector spaces with different coordinate systems. This means that the temporal models from the time intervals must be aligned before using them to

[3]More details on ERDE metric are described in Subsection 4.1

compare representations, which comes at computational and performance costs. Static-based TWEs considered different *alignment* strategies. Most of them assume that words at nearby periods are highly similar, applying strategies to minimize the distance between contiguous vector space models (VSM).

A common way of approaching the learning of TWE is to divide data into time slices and learn static embeddings of each one independently, and then proceed to the alignment of the models [55, 23, 56, 57, 58, 59]. Following this path, Hamilton et al. [55] computed the artificial rotation between nearby spaces and used it to calculate real semantic shifts; while others like Kim et al. [56] addressed the alignment through neural network initialization. In general, such strategies struggle to generate optimal solutions due to the considerable reduction in training data. Other recent works train all the temporal vectors concurrently, enforcing all of them inside a common space [58, 59, 57]. Yao et al. [58] applied regularization terms to smooth embedding changes across time. Rudolph and Blei [57] and Bamler and Mandt [59] cast the embeddings in probabilistic frameworks that also smooth the vectors' trajectory. These strategies produce better results in smaller datasets since they can use all of the data in training but are less scalable and efficient. Temporal Word Embeddings with a Compass (TWEC) [24] is a recent model that does not require explicit alignment between different time intervals, which alleviates alignment constraints. Instead, TWEC uses atemporal vectors as a reference (*compass*) when training the temporal representations. In this way, the representations share a unique coordinate system.

Recently, research efforts are centred on detecting language change in contextualized word embeddings [60, 61, 62, 25]. Pre-trained transformers made it possible to contextualize the training of temporal embeddings for shorter time slices, e.g., single years. However, it is worth noting that contextualized embeddings perform worse than static embeddings for semantic change detection tasks [63] so far. In this regard, Dynamic Contextualized Word Embeddings (DCWE) [25] is a recent approach that combines the strengths of contextualized word embeddings with the flexibility of dynamic word embeddings. DCWE jointly models temporal and social information using pre-trained language models across time intervals. The authors demonstrated that considering time information can even improve downstream tasks such as sentiment analysis.

3. Proposal

We based our proposal on the idea that individuals developing depressive symptoms will eventually change their language usage. They express new feelings, thoughts, and emotions, which will induce changes in their writings. Consequently, depressed individuals will vary the use of words, and many words will appear in previously unseen contexts. We designed a framework to analyze the variations in the individual's use of language over time. Our method locates word shifts from one time step to another and quantifies these meaning changes. Next, we will introduce the overall view of the framework for later on going on the different implementations that we tested.

Our framework calculates two types of word embeddings: reference and dynamic. Reference embeddings are static and capture standard and neutral use of words. Contrarily, the dynamic embeddings represent the meaning of words for a user in each specific time interval, i.e., we

compute a different dynamic vector space for each chunk and user. As a result, we can quantify the magnitude of word movements over different time slices. We obtain these movements by computing the distance between the dynamic word embeddings and the reference vectors. Then, the movements of words from chunk to chunk are the features considered during the training and inference classification phases. The general framework is divided into three main phases: (i) construction of reference embeddings, (ii) construction of dynamic embeddings and (iii) magnitude calculation of word movements. The first two phases depend on the selected temporal model. The last one (computation of word movements) is model-independent and remains the same regardless of the variant.

(i) First, we construct the global reference space, R. This global space is atemporal, considering the full training corpus. We train a model computing a reference vector for each word w present in the corpus vocabulary V with the whole training corpus.

(ii) After computing the reference space, we consider the n temporal chunks for a user. We define C^{t_i} as the temporal spaces of each chunk, with $0 \leq i \leq n$. Each point in $TC^{t_i}(w)$ corresponds to the usage of the word w at time t_i. The training process of the reference and temporal spaces varies depending on the embedding model considered. This work presents two variants by using static and contextualized embedding architectures.

(iii) Finally, we quantify the magnitude of the word shifts that occurred over time. Our framework obtains the word shifts on the user's writings of each chunk TC^{t_i} by computing the distance of each word from the chunk t_i to that same word in the *reference vector space* R. We define these *word shifts* $\delta_w^{t_i}$ as the movement of the word w at time t_i. The word shifts ($\delta_w^{t_i}$) are from now on called deltas. Bigger values of deltas will represent larger deviations from the reference. If a word is not present in a specific chunk TC^{t_i}, we consider that its meaning has not changed, and its delta has zero value. To analyze variations in the individual's use of language, we construct a matrix of deltas based on their writings. We defined this user matrix as Δ_{u_r}. Thus, for an user u_r, we construct the matrix as equation 1:

$$\Delta_{u_r} = \begin{array}{c} \\ t_1 \\ \vdots \\ t_i \\ \vdots \\ t_n \end{array} \begin{array}{c} \begin{array}{cccccc} w_1 & \cdots & w_k & \cdots & w_m \end{array} \\ \begin{pmatrix} \delta_{w_1}^{t_1} & \cdots\cdots & \delta_{w_k}^{t_1} & \cdots\cdots & \delta_{w_m}^{t_1} \\ \vdots & & \vdots & & \vdots \\ \delta_{w_1}^{t_i} & \cdots\cdots & \delta_{w_k}^{t_i} & \cdots\cdots & \delta_{w_m}^{t_i} \\ \vdots & & \vdots & & \vdots \\ \delta_{w_1}^{t_n} & \cdots\cdots & \delta_{w_k}^{t_n} & \cdots\cdots & \delta_{w_m}^{t_n} \end{pmatrix} \end{array} \tag{1}$$

where rows represent the n temporal chunks of that user, and columns the movements for each word w_k present in the vocabulary, $w_k \in V$. Each row of the user matrix represents the word shifts for a different chunk. For instance, the row t_i ($\Delta_{u_r}[t_i]$) captures all the word shifts that happened for the individual during the chunk i.

We addressed the problem as a binary classification (non-depressed vs depressed), and trained the classifiers at chunk-level. Therefore, each row of the users' matrix is a different data sample. In the training phase, each row is labelled as 0 or 1 depending on whether the training user is negative or positive. In the inference phase, we feed the trained classifiers one user chunk at a time in chronological order. If the classifier's decision over that chunk is negative (non-depressed), we continue processing the subsequent user's chunks. If the decision is positive,

we emit an alert for the user. However, if our classifiers reach the last temporal chunk and the decision is still negative, we classify the test user as non-depressed. We give more details of the classification process in the subsection 4.2.

3.1. Construction of Reference and Temporal Spaces

This subsection describes the different choices for building the word representations. We implemented two variants based on the adaptation of existing temporal models. We explain how we constructed the reference and temporal vector spaces. After computing those spaces, the subsequent calculation of words shifts remains the same in both variants.

3.1.1. Temporal Word Embeddings with a Compass (TWEC)

First, we adapted a temporal model based on word2vec, known as the TWEC method [24]. TWEC proposes the use of atemporal vectors as references, i.e. as a *compass*, when training the specific representations to a given time interval. This model fits very well with our framework since we can take advantage of the compass construction and use it as our reference vector space. TWEC approach is based on the premise that words that have changed over time appear in contexts with words whose meaning changed slightly. For that, authors exploit the word2vec architecture, which employs a dual representation of words (input and output embeddings). TWEC keeps output embeddings frozen across time and trains the temporal slices updating only their input embeddings. The frozen layer acts as an atemporal compass, allowing all the temporal models to share the same coordinate system. In the original study, the authors trained the compass with the whole experimental datasets. Then, they divided the collections into temporal chunks and used the compass to generate the temporal spaces. However, due to the eRisk task's nature, at inference time, we do not have access to the entire history of the writings of a test user. That is, proceeding as in the original TWEC application would be cheating on the eRisk task. In this case, our model needs to process individual chunks sequentially and make early alerts using only the information related to that chunk. Therefore, we needed to modify TWEC to leverage it for our task slightly.

Our adaptation of TWEC is graphically depicted in Figure 1. First, we construct the input and output weight matrices, what we define as I and R, by training the word2vec CBOW architecture on the entire training collection. We ignore the resulting I matrix and consider the output matrix R as our reference space for temporal chunks. Therefore, the reference space model captures a general use of language by the Reddit users (including depressed and control users). After R is calculated, we can start computing all the temporal chunk models $TC_{u_i}^{t^k}$ associated to each user u_i and to a specific time t^k. We train every temporal model using the user's writings for that chunk as input. As output, we keep the R embeddings frozen to ensure that all the temporal models share the same coordinate system. This process remains the same when calculating the temporal spaces for training and test users for the subsequent classification task.

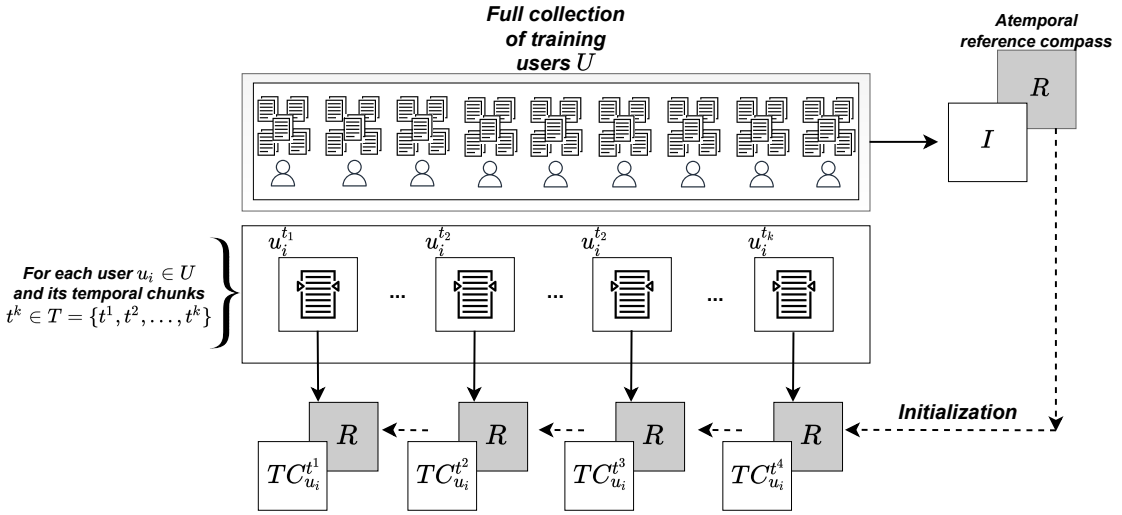

Figure 1: TWEC adaptation for our method. The temporal chunk embeddings $TC_{u_i}^{t_k}$ are trained over each chunk slice. Temporal embeddings use the atemporal compass acting as reference R.

3.1.2. Dynamic Contextualized Word Embeddings (DCWE)

We experimented with another TWE variant to calculate the reference and temporal models using contextualized word embeddings. In this case, we adapted the Dynamic Contextualized Word Embeddings (DCWE) [25] proposal to compute the vector spaces. The idea is to exploit the transformers' architecture for generating context-based embeddings. In this model, we can distinguish two main components: (a) Dynamic module and (b) Contextualizer, illustrated on Figure 2. Words are first mapped to dynamic representations and then are contextualized. For both components, we used the standard $\text{BERT}_{\text{BASE}}$ language model, with 12 attention layers and a hidden layer with 768 dimensions. The resulting word embeddings are called e_j^k for word k and time step j. The dynamic component is inspired by temporal models proposed in early works in the literature. Authors model a vector offset that depends on word and time. This offset (o_j^k) is generated by a time-specific feed-forward network (FFN). To model the temporal drift, authors impose a random walk prior over o_j^k. The random walk enforces smooth changes of the dynamic embeddings over subsequent times. Consequently, the distance between words that are far apart in time will be greater. The offset is summed to the static embeddings $\tilde{e}_j^k$, which are initialized using BERT pre-trained embeddings.

The concatenation of $\tilde{e}_j^k$ and o_j^k embeddings is the dynamic component of the final embedding. Then, these dynamic embeddings are contextualized using BERT (component b in the Figure). The contextualizer applies masked language modelling (MLM) as the training objective. The outputs of both components are the final *dynamic contextualized* embeddings. We note that, due to the maximum length limit of BERT, we need to split the chunks into individual posts. This results in different dynamic word embeddings for each word appearing in the writings.

With DWEC, we produce dynamic contextualized embeddings. However, we still miss the reference space vector R to compute the temporal word shifts. To create R, we have taken advantage of the previously static embeddings. We believe that the BERT embeddings, trained

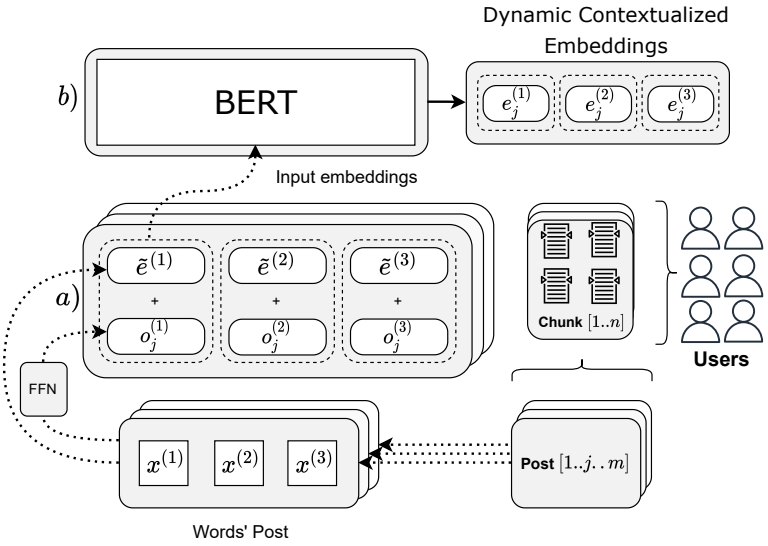

Figure 2: DCWE implementation of our variant. Due to the length max limit of BERT inputs (512 tokens), the dynamic embeddings are trained at post-level. This implies that we now have a different word representation for each time a word appears in a post. For more details of the components, please see [25].

in large collections of documents, would work as a good reference space in our solution. Finally, we can compare the temporal word shifts from the reference embeddings. For clarification purposes, this method is illustrated in Figure 3. As previously mentioned, we perform the whole process at the post level. Therefore, we calculate the deltas also at the post level. Thus, to quantify the word shifts for each chunk, we averaged all the distances of the words' used in that time interval [4].

4. Experiments

This section covers the experimental settings and evaluation of our proposed approaches. First, we briefly describe the collections used to train and test our classifiers. We also introduce the evaluation metrics that consider both the effectiveness and the delay in making the decisions. Then, subsection 4.2 presents the implementation details of our experiments. Finally, we discuss the results and compare our performance against state-of-the-art baselines.

4.1. Dataset and evaluation metrics

The collections used in this work correspond with the eRisk task *early risk detection of depression* (EDD) from 2017 and 2018. This task is focused on sequentially processing the publications posted by users on Reddit[5]. There are two classes of users, *depressed* and *non-depressed*. Table

[4]If a word has not occurred in any of the posts' chunks, the shift value for that word is zero (0)
[5]https://www.reddit.com/

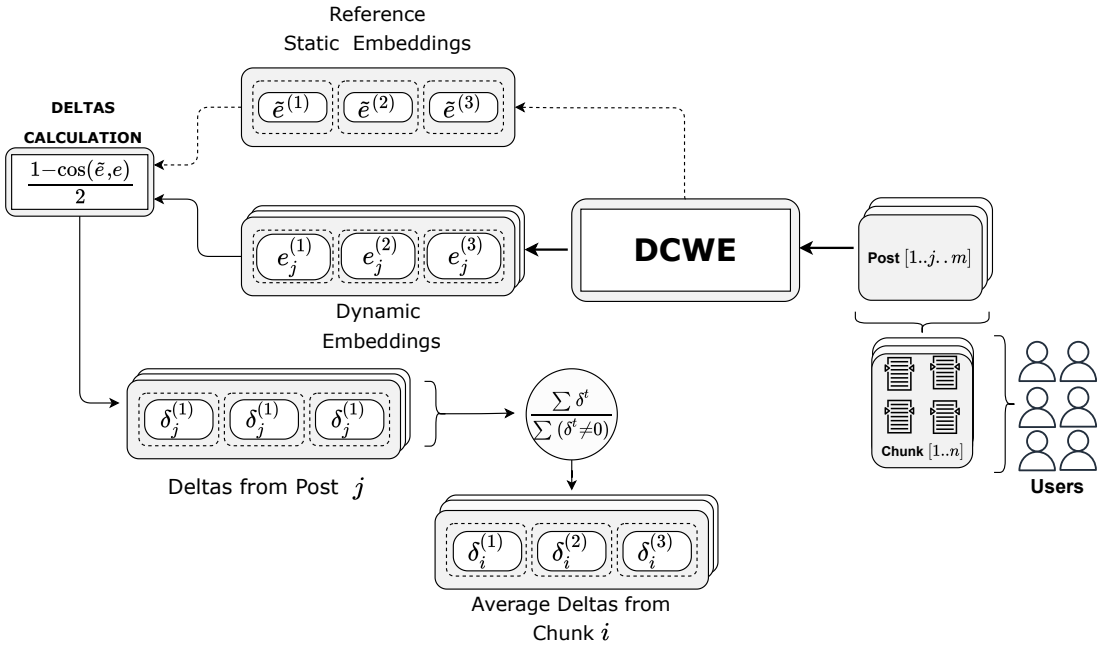

Figure 3: Calculation of the word shifts of each chunk after the DCWE model is trained. As we have a different embedding for each word occurence in the writings, we average all its embeddings to quantify each word shift.

1 summarizes the details of each collection. Note that both datasets are highly unbalanced, as there are many more control users. User's writings are chronologically sorted into ten chunks. Thus, each chunk contains 10% of the subject's history. Alerts of the user developing depression must be emitted as early as possible, deciding each user is depressed or not based on the sequential chunks of information. For this purpose, classifiers are given one chunk at a time, and then they have to label the user as depressed, non-depressed, or they need more chunks before making a decision.

Table 1
Statistics of eRisk 2017 and 2018 collections for the depression estimation task

Edition	2017		2018	
	Train	Test	Train	Test
Depressed	83	52	135	79
Non-depressed	403	349	752	741
Total	486	401	887	820

Regarding the evaluation metrics, organizers proposed standard classification metrics such as F_1-measure, precision (p) and recall (r). Additionally, as the main objective of the task is to emit alerts as early as possible, a novel time-aware measure was defined by Losada and Crestani [31]. This new metric, called *Early Risk Detection Error* (ERDE), considers both the correctness of the

predictions and the delay it took to emit the alert. ERDE is defined by:

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d \text{ is False Positive (FP)} \\ c_{fn} & \text{if } d \text{ is False Negative (FN)} \\ lc_0(k) \cdot c_{tp} & \text{if } d \text{ is True Positive (TP)} \\ 0 & \text{if } d \text{ is True Negative (TN)} \end{cases}$$

d is the binary decision taken by the systems with delay k for the user. The cost function $lc_0(k)$ is a monotonically increasing function of k and o, the penalization threshold:

$$lc_0(k) = 1 - \frac{1}{1 + e^{k-o}} \tag{2}$$

The delay is measured by counting the number of writings (k) processed before deciding. Particularly, organizers proposed the use of *ERDE₅* and *ERDE₅₀*, where 5 and 50 represent the number of posts from which the metric penalizes more quickly. Finally, c_{fn} and c_{tp} costs are set to 1, and c_{fp} cost is the proportion of depressed subjects from the total subjects in the test set. Being error measures, lower values of ERDE mean better performance.

4.2. Implementation details

4.2.1. Filtering process

When calculating the word deviations, we have experimented with different lexicons filtering out terms by their total term frequency in the corpus. The total unique words from the 2017 and 2018 corpora are 42588 and 62455, respectively. With this in mind, we have considered three different settings to see how the words' frequency might affect the performance of our solutions. (i) Filter out the 100 most frequent words, (ii) filter out the 100 most frequent words and only use the next 10000 most frequent terms [64] and (iii) consider all the terms in the corpus. The selection of terms was made only over the training data, unseen terms on test data were discarded. For selecting the best filtering strategy, we cross-validated it on the training split while optimizing for the $ERDE_{50}$ metric. The results reported in this paper use the first configuration, as it is the setting that obtained the best values in the validation stage. This seems to point out that temporal shifts from words with a a high term frequency may add noise to the process, while keeping low frequent terms can improve the classification. We leave for future work a thorough comparison between the impact of the term frequency and term selection methods on the performance of our temporal solutions.

4.2.2. Classification process

We perform a binary classification task (depressed vs non-depressed) by considering as features only the word shifts at each chunk. In this process, we used a standard classifier, Support Vector Machine (SVM) of Python sklearn. We decided to experiment with SVM's motivated by its remarkable results in the eRisk tasks compared to other classification techniques. We experimented with different kernels of SVM during the validation stage. Parameters were optimized to minimize the $ERDE_{50}$ measure. We applied a grid search strategy (ranging from

0.5 to 3) for the parameter C, and class weights parameter (ranging from 1 to 23). The SVM with RBF kernel achieved the best results for this metric. Therefore, the results reported by our approaches use this setting. We leave for future work to evaluate other classifiers.

4.3. Results and discussion

The objective of this paper was to evaluate the adaptation of temporal embedding models to early detection scenarios. For that, we compared our results against the best participants and state-of-the-art methods for the eRisk early task in 2017 and 2018. Table 2 shows the baselines along with our proposed methods (**ΔTWEC_SVM** and **ΔDCWE_SVM**).

Table 2
Results of our model and the best baselines for each metric in eRisk 2017 and 2018. The best results are bolded. For ERDE metrics, the lower the better.

Edition	Model	$ERDE_5$	$ERDE_{50}$	*F1*
2017	Team FHDOB	12.70	10.39	0.55
	Team UNSLA	13.66	9.68	0.59
	Team FHDOA	12.82	9.69	**0.64**
	SS3 [45]	12.70	**7.72**	0.54
	ΔTWEC_SVM	12.83	9.4	0.56
	ΔDCWE_SVM	**12.44**	8.61	0.51
2018	Team FHDOB	9.50	6.44	**0.64**
	Team UNSLA	**8.78**	7.39	0.32
	ModComb [49]	9.52	**6.12**	0.51
	ΔTWEC_SVM	9.99	6.99	0.53
	ΔDCWE_SVM	9.12	6.34	0.49

In 2017, there were 30 submissions from 8 international participants. All the participants' results' are detailed in [18]. We report the best performers from the participants in $ERDE_5$, $ERDE_{50}$ and F_1. We also include the SS3 method [45], a recent text classification framework that obtained state-of-the-art results in the task. We did not report values for this model in 2018 as the authors did not provide those numbers. *FHDOB* and *FHDOA* [46] proposals are based on robust systems with a wide set of features, ranging from specific lexicons (for instance, terms related to antidepressants) to automatically learned features extracted with the use of neural networks. *FHDOB* focuses on document vectorization with *doc2vec* [65] and ranked first in $ERDE_5$. *FHDOA* uses an ensemble of *Bag of Words* with term weightings and different n-grams, being the best solution in F_1. Finally, *UNSLA* analyses the variation of vocabulary in the chunks [66] using semantic representations of documents and were the best solution in $ERDE_{50}$. In this edition, our two approaches would achieve the best result among participants according to $ERDE_{50}$. Moreover, the ΔDCWE_SVM also ranks first in $ERDE_5$ even including the SS3 model. However, we observe that our performance remains slightly below SS3 in both error metrics. Finally, our $F1$ performance is not so good but was still above the average of all participants (0.39). This drop regarding $F1$ is reasonable since we tuned out parameters to

prioritise early alerts rather than accuracy.

In 2018, there were 45 systems from 11 different institutions. Again, the participants' performance can be consulted in the CLEF official working notes of 2018 [19]. Coincidentally, the best performers for each metric were from the same research groups as in 2017. *FHDOB*'s incremental solution disregarded almost all the metadata features [67], and obtained the best results in $ERDE_{50}$ and $F1$. The *UNSLA* method [48] improved their previous method by providing a flexible new approach that allows varying the number of chunks considered in the representation according to the level of urgency required. Additionally, Ramiandrisoa et al. [49] recently proposed a machine learning model, *ModComb*, that combines (*a*) state-of-the-art features extracted from user's writings and (*b*) text representation using word embeddings. Some of the *ModComb* features were specially designed for depression and outperformed all the existing methods in $ERDE_{50}$. Again, the authors did not provide evaluation results of their method in 2017, so we could not report them. In this edition, ΔDCWE_SVM outperformed all eRisk participants in $ERDE_{50}$ and was only surpassed by *UNSLA* in $ERDE_5$. However, the values reported by *ModComb* remain slightly better than us in terms of $ERDE_{50}$. We also note that the ΔTWEC_SVM approach ranks among the top-performing methods, including our DCWE variant. The $F1$ performance follows the same trend as in 2017, as our approaches are still better than above average (0.41).

We can conclude from the overall results that capturing changes in individuas' use of language over time show remarkable performance in early detection scenarios. Our proposed approaches obtained competitive results compared to the state-of-the-art, especially considering the time-aware error measures. Even in some cases, we outperformed all the reported values for the task. Additionally, we have to stress that, in contrast to most previous solutions, (i) our methods do not apply an elaborated set of hand-crafted and specifically designed combination of features, such as the exploit of depression or emotion lexicons seen at prior works. (ii) The decisions do not rely on complex models, ensembles of different classifiers, etc. In contrast to most proposed approaches, our proposal is easily adaptable to other domains and could even be combined with other specific features without any additional effort.

5. Conclusions and Future Work

This paper explored the potential of language evolution in early depression detection using publications from social media. A basic idea inspired our work: when people develop a depressive condition, they regularly change their use of language, i.e., the way they communicate with others. We proposed and evaluated a novel early detection framework to assess this assumption. This framework adapts temporal word representations to calculate the magnitude of word movements over time. Particularly, we implemented two temporal techniques based on (i) the use of static embedding models and (ii) contextualized embedding models. We evaluated both approaches on two datasets provided at eRisk 2017 and 2018 in CLEF international forum. Our results show how useful it can be to consider the evolution of words as features to detect early traces of depression. By considering only word shifts as features, we ranked among the top state-of-the-art models. Another interesting aspect of our framework is that we do not rely on depression-specific features or elaborated decision mechanisms to emit the alerts.

The flexibility of our framework opens many research lines for future work. First, we plan to extend the training of the reference space model over a larger number of users to increase its stability. In this regard, we are interested in using other datasets to confirm our results, considering other social platforms and languages. Additionally, this adaptability also allows us to investigate the commonalities in terms of words' evolution with other similar diseases such as anorexia, self-harm or suicidal ideation. Finally, we will explore ways to improve the performance of our solutions. We will analyze other classification algorithms and feature selection methods better suited for handling large scale data and noisy features.

References

[1] S. Saxena, M. Funk, D. Chisholm, Comprehensive mental health action plan 2013–2020, EMHJ-Eastern Mediterranean Health Journal 21 (2015) 461–463.

[2] R. C. Kessler, K. A. McGonagle, S. Zhao, C. B. Nelson, M. Hughes, S. Eshleman, H.-U. Wittchen, K. S. Kendler, Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey, Archives of general psychiatry 51 (1994) 8–19.

[3] S. I. Hay, A. A. Abajobir, Abate, et al., Global, regional, and national disability-adjusted life-years (dalys) for 333 diseases and injuries and healthy life expectancy (hale) for 195 countries and territories, 1990–2016: a systematic analysis for the global burden of disease study 2016, The Lancet 390 (2017) 1260–1344.

[4] J. Bueno-Notivol, P. Gracia-García, B. Olaya, I. Lasheras, R. López-Antón, J. Santabárbara, Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies, International Journal of Clinical and Health Psychology 21 (2021) 100196.

[5] R. Mojtabai, M. Olfson, B. Han, National trends in the prevalence and treatment of depression in adolescents and young adults, Pediatrics 138 (2016).

[6] P. Sobocki, B. Jönsson, J. Angst, C. Rehnberg, Cost of depression in europe., Journal of Mental Health Policy and Economics (2006).

[7] A. Barak, J. M. Grohol, Current and future trends in internet-supported mental health interventions, Journal of Technology in Human Services 29 (2011) 155–196.

[8] A. Halfin, Depression: the benefits of early and appropriate treatment, American Journal of Managed Care 13 (2007) S92.

[9] A. Picardi, I. Lega, L. Tarsitani, M. Caredda, G. Matteucci, M. Zerella, R. Miglio, A. Gigantesco, M. Cerbo, A. Gaddini, et al., A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care, Journal of affective disorders 198 (2016) 96–101.

[10] J. W. Pennebaker, M. R. Mehl, K. G. Niederhoffer, Psychological aspects of natural language use: Our words, our selves, Annual review of psychology 54 (2003) 547–577.

[11] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of LIWC2015, Technical Report, 2015.

[12] R. A. Calvo, D. N. Milne, M. S. Hussain, H. Christensen, Natural language processing in mental health applications using non-clinical texts, Natural Language Engineering 23 (2017) 649–685.

[13] N. Colineau, C. Paris, Talking about your health to strangers: understanding the use of online social networks by patients, New review of hypermedia and multimedia 16 (2010) 141–160.
[14] J. Suler, The online disinhibition effect, Cyberpsychology & behavior 7 (2004) 321–326.
[15] C. M. McHugh, A. Corderoy, C. J. Ryan, I. B. Hickie, M. M. Large, Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value, BJPsych open 5 (2019).
[16] R. N. Golden, C. Weiland, F. Peterson, The truth about illness and disease, Infobase Publishing, 2009.
[17] L. Manikonda, M. De Choudhury, Modeling and understanding visual attributes of mental health disclosures in social media, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 170–181.
[18] D. E. Losada, F. Crestani, J. Parapar, eRisk 2017: CLEF lab on early risk prediction on the internet: experimental foundations, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2017, pp. 346–360.
[19] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk: early risk prediction on the internet, in: International conference of the cross-language evaluation forum for european languages, Springer, 2018, pp. 343–361.
[20] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2019 early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 340–357.
[21] D. E. Losada, F. Crestani, J. Parapar, eRisk 2020: Self-harm and depression challenges, in: European Conference on Information Retrieval, Springer, 2020, pp. 557–563.
[22] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, 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, Virtual Event, September 21-24, 2021, Proceedings, Springer International Publishing, 2021, pp. 324–344.
[23] V. Kulkarni, R. Al-Rfou, B. Perozzi, S. Skiena, Statistically significant detection of linguistic change, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 625–635.
[24] V. Di Carlo, F. Bianchi, M. Palmonari, Training temporal word embeddings with a compass, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 6326–6334.
[25] V. Hofmann, J. Pierrehumbert, H. Schütze, Dynamic contextualized word embeddings, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021.
[26] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[27] D. J. Harper, Discourse analysis and'mental health', Journal of Mental Health 4 (1995) 347–358.
[28] M. E. Jensen, E. A. Pease, K. Lambert, D. R. Hickman, O. Robinson, K. T. McCoy, J. K. Barut, K. M. Musker, D. Olive, C. Noll, et al., Championing person-first language: a call to psychiatric mental health nurses, Journal of the American Psychiatric Nurses Association 19 (2013) 146–151.
[29] T. Althoff, K. Clark, J. Leskovec, Large-scale analysis of counseling conversations: An application of natural language processing to mental health, Transactions of the Association for Computational Linguistics 4 (2016) 463–476.
[30] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses, in: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2015, pp. 1–10.
[31] D. E. Losada, F. Crestani, A test collection for research on depression and language use (2016) 28–39.
[32] S. MacAvaney, B. Desmet, A. Cohan, L. Soldaini, A. Yates, A. Zirikly, N. Goharian, Rsdd-time: Temporal annotation of self-reported mental health diagnoses, arXiv preprint arXiv:1806.07916 (2018).
[33] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, N. Goharian, Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions, arXiv preprint arXiv:1806.05258 (2018).
[34] D. Mowery, H. Smith, T. Cheney, G. Stoddard, G. Coppersmith, C. Bryan, M. Conway, Understanding depressive symptoms and psychosocial stressors on twitter: a corpus-based study, Journal of medical Internet research 19 (2017) e6895.
[35] M. Mitchell, K. Hollingshead, G. Coppersmith, Quantifying the language of schizophrenia in social media, in: Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality, 2015, pp. 11–20.
[36] B. Desmet, V. Hoste, Online suicide prevention through optimised text classification, Information Sciences 439 (2018) 61–78.
[37] M. De Choudhury, E. Kiciman, The language of social support in social media and its effect on suicidal ideation risk, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 11, 2017.
[38] T. Wang, M. Brede, A. Ianni, E. Mentzakis, Detecting and characterizing eating-disorder communities on social media, in: Proceedings of the Tenth ACM International conference on web search and data mining, 2017, pp. 91–100.
[39] X. Chen, M. D. Sykora, T. W. Jackson, S. Elayan, What about mood swings: Identifying depression on twitter with temporal measures of emotions, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 1653–1660.
[40] M. A. Moreno, L. A. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, T. Becker, Feeling bad on facebook: Depression disclosures by college students on a social networking site, Depression and anxiety 28 (2011) 447–455.
[41] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Seventh international AAAI conference on weblogs and social media, 2013.
[42] M. De Choudhury, S. Counts, E. Horvitz, Social media as a measurement tool of depression

in populations, in: Proceedings of the 5th annual ACM web science conference, 2013, pp. 47–56.
[43] M. De Choudhury, S. De, Mental health discourse on reddit: Self-disclosure, social support, and anonymity, in: Eighth international AAAI conference on weblogs and social media, 2014.
[44] R. M. Ortega-Mendoza, D. I. Hernández-Farías, M. Montes-y Gómez, L. Villaseñor-Pineda, Revealing traces of depression through personal statements analysis in social media, Artificial Intelligence in Medicine 123 (2022) 102202.
[45] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197.
[46] M. Trotzek, S. Koitka, C. Friedrich, Linguistic metadata augmented classifiers at the CLEF 2017 task for early detection of depression, 2017.
[47] M. P. Villegas, D. G. Funez, M. J. G. Ucelay, L. C. Cagnina, M. L. Errecalde, LIDIC-UNSL's participation at eRisk 2017: Pilot task on early detection of depression., in: CLEF (Working Notes), 2017.
[48] D. G. Funez, M. J. G. Ucelay, M. P. Villegas, S. Burdisso, L. C. Cagnina, M. Montes-y Gómez, M. Errecalde, UNSL's participation at eRisk 2018 lab., in: CLEF (Working Notes), 2018.
[49] F. Ramiandrisoa, J. Mothe, Early detection of depression and anorexia from social media: A machine learning approach, in: Circle 2020, volume 2621, 2020.
[50] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).
[51] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[52] J. Sarzynska-Wawer, A. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz, L. Okruszek, Detecting formal thought disorder by deep contextualized word representations, Psychiatry Research 304 (2021) 114135.
[53] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[54] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.
[55] W. L. Hamilton, J. Leskovec, D. Jurafsky, Diachronic word embeddings reveal statistical laws of semantic change, arXiv preprint arXiv:1605.09096 (2016).
[56] Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, S. Petrov, Temporal analysis of language through neural language models, in: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 61–65. URL: https://aclanthology.org/W14-2517. doi:10.3115/v1/W14-2517.
[57] M. Rudolph, D. Blei, Dynamic embeddings for language evolution, in: Proceedings of the 2018 World Wide Web Conference, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, p. 1003–1011. URL:

https://doi.org/10.1145/3178876.3185999. doi:10.1145/3178876.3185999.
[58] Z. Yao, Y. Sun, W. Ding, N. Rao, H. Xiong, Dynamic word embeddings for evolving semantic discovery, 2018, pp. 673–681. doi:10.1145/3159652.3159703.
[59] R. Bamler, S. Mandt, Dynamic word embeddings, in: International conference on Machine learning, PMLR, 2017, pp. 380–389.
[60] M. Giulianelli, M. Del Tredici, R. Fernández, Analysing lexical semantic change with contextualised word representations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3960–3973. URL: https://aclanthology.org/2020.acl-main.365. doi:10.18653/v1/2020.acl-main.365.
[61] R. Hu, S. Li, S. Liang, Diachronic sense modeling with deep contextualized word embeddings: An ecological view, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3899–3908. URL: https://aclanthology.org/P19-1379. doi:10.18653/v1/P19-1379.
[62] M. Martinc, P. K. Novak, S. Pollak, Leveraging contextual embeddings for detecting diachronic semantic shift, 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, European Language Resources Association, 2020, pp. 4811–4819. URL: https://www.aclweb.org/anthology/2020.lrec-1.592/.
[63] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, N. Tahmasebi, Semeval-2020 task 1: Unsupervised lexical semantic change detection, arXiv preprint arXiv:2007.11464 (2020).
[64] H. P. Luhn, The automatic creation of literature abstracts, IBM Journal of research and development 2 (1958) 159–165.
[65] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.
[66] M. L. Errecalde, M. P. Villegas, D. G. Funez, M. J. G. Ucelay, L. C. Cagnina, Temporal variation of terms as concept space for early risk prediction., in: CLEF (Working Notes), 2017.
[67] M. Trotzek, S. Koitka, C. M. Friedrich, Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia., in: CLEF (Working Notes), 2018.