Contextualized Diachronic Word Representations

Ganesh Jawahar **Djamé Seddah**
Inria
{firstname.lastname}@inria.fr

Abstract

Diachronic word embeddings play a key role in capturing interesting patterns about how language evolves over time. Most of the existing work focuses on studying corpora spanning across several decades, which is understandably still not a possibility when working on social media-based user-generated content. In this work, we address the problem of studying semantic changes in a large Twitter corpus collected over five years, a much shorter period than what is usually the norm in diachronic studies.

We devise a novel attentional model, based on Bernoulli word embeddings, that are conditioned on contextual extra-linguistic (social) features such as network, spatial and socioeconomic variables, which are associated with Twitter users, as well as topic-based features. We posit that these social features provide an inductive bias that helps our model to overcome the narrow time-span regime problem. Our extensive experiments reveal that our proposed model is able to capture subtle semantic shifts without being biased towards frequency cues and also works well when certain contextual features are absent. Our model fits the data better than current state-of-the-art dynamic word embedding models and therefore is a promising tool to study diachronic semantic changes over small time periods.

1 Introduction

Natural language changes over time due to a wide range of linguistic, psychological, sociocultural and encyclopedic causes (Blank and Koch, 1999; Grzega and Schoener, 2007). Studying the semantic change of a word helps us understand more about the human language and build temporally aware models, that are especially complementary to the work done in the digital humanities and historical linguistics. Recently, diachronic word embeddings based on distributional hypothesis (Harris, 1954) have been used to automatically study semantic changes in a data-driven fashion from large corpora (Kim et al., 2014; Hamilton et al., 2016; Rudolph and Blei, 2018). We refer the reader to Kutuzov et al. (2018) who survey the recent methods in this field and establishes the challenges that lie ahead.

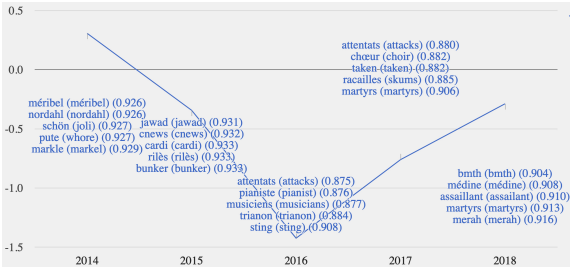


Figure 1: The diachronic embedding computed by our proposed model for the word 'BATACLAN' reveals how the term's usage changed over the years. We list the most similar five words (with English translation in paranthesis) in each year by cosine similarity. The y-axis corresponds to "meaning", a one dimensional PCA projection of the embeddings.

Currently, we find the literature on this problem to be focused on English corpora, spanning across several decades. This has not only created a gap in extending the diachronic word embeddings for a wider scope of languages, but also to datasets spanning across few successive years which are common in digital humanities and social sciences. In this work, we study French text from Twitter collected over just five years, which provides a challenging platform to build models that can capture semantic drifts in a noisy, subtly evolving language corpus.

Figure 1 shows an instance of the evolution of the word 'Bataclan' (a theatre in Paris that was at-

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 35–47
Florence, Italy, August 2, 2019.

tacked by terrorists on November 2015) from the French corpus. It also shows that such embedding representations mostly capture the dominant sense of a word when used in synchrony and can therefore only reflect the evolution of the dominant sense when used diachronically, yet leaving open the question of whether small, subtle changes can be captured (Tahmasebi et al., 2018).

We hypothesize that the current state-of-the-art models lack inductive biases to fit data accurately in this setting. We build on the observation by Jurafsky (2018) that *"it's important to consider who produced the language, in what context, for what purpose, and make sure that the models are fit to the data"*. Hence, we propose a novel model extending on Dynamic Bernoulli word Embeddings (Rudolph and Blei, 2018) (DBE) which exploits the inductive bias by conditioning on a number of contextualized features such as network, spatial and socio-economic variables, which are associated with Twitter users, as well as topic-based features.

We perform qualitative studies and show that our model can: (i) accurately capture the subtle changes caused due to cultural drifts, (ii) learn a smooth trajectory of word evolution despite exploiting various inductive biases. Our quantitative studies illustrate that our model can: (i) capture better semantic properties, (ii) be less sensitive to frequency cues compared to DBE model, (iii) act as better features for 2 out of 4 tweet classification tasks. Through an ablation study, we find in addition that our model can: (iv) work with a reduced set of contextualized features, (v) follow the test of law of prototypicality (Dubossarsky et al., 2015). In sum, we believe our model is a promising tool to study diachronic semantic changes over small time periods. [1]

Our main contributions are as follows:

- Our work is the first to study diachronic word embeddings for tweets from French language to the best of our knowledge. Unlike previous works, we consider dataset from a narrow time horizon (five years).
- We propose a novel, attentional, diachronic word embedding model that derives inductive biases from several contextualized, socio-demographic, features to fit the data accurately.
- Our work is also the first to estimate the usefulness of the diachronic word embeddings for downstream task like tweet classification.

2 Related Work

Kim et al. (2014) introduced prediction-based word embedding models to track semantic shifts across time. They extended SkipGram model with Negative Sampling (SGNS) (Mikolov et al., 2013) by training a model on current year after initializing the word embeddings from trained model of previous year. This initialization ensures the word vectors across time slices are grounded in same semantic space. Kulkarni et al. (2015) and Hamilton et al. (2016) utilize ad hoc alignment techniques like orthogonal Procrustes transformations to map successive model pairs together. These approaches have an impractical demand of having enough data in each time slice to learn high quality embeddings.

The work done by Bamler and Mandt (2017), Yao et al. (2018) and Rudolph and Blei (2018) proposed to learn word embeddings across all time periods jointly along with their alignment in a single step. Rudolph and Blei (2018) represent word embeddings as sequential latent variables, naturally accommodating for time slices with sparse data and assuring word embeddings are grounded across time. Our proposed model builds upon this work to condition on several inductive biases, using contextual extra-linguistic (social) and topic-based features, to accurately fit dataset from a narrow time horizon.

3 Contextualized Features

Natural language text is inherently contextual, depending on the author, the period and the intended purpose (Jurafsky, 2018). For instance, features based on authors' demography although incomplete can explain some of the variance in the text (Garten et al., 2019). While diachronic word embeddings' ability to capture semantic shifts is interesting because of its flexibility, we postulate that there is a need to capture contextualized information about tweets such as the characteristics of their authors (including spatial, network, socio-economic, interested topics) and meta-information such as their topic. To extract features, we make use of the largest French Twitter corpus to date proposed in Abitbol et al. (2018). In this section we will describe the set of contextualized feature

[1] Code to reproduce our experiments is publicly accessible at https://github.com/ganeshjawahar/social_word_emb

we propose to inject to our diachronic word embedding model (see Section 4).

3.1 Spatial

Users from similar geographical areas tend to share similar properties in terms of word usage and language idiosyncrasies. Among others, Hovy and Purschke (2018) for German and Abitbol et al. (2018) for French, confirmed regional variations in geolocated users' content in social media. The latter work found the southern part of France to use a more standard language than the northern part. To exploit these geographic variations, we identify geolocated users ($\sim 100K$) and associate each of them to their respective region (out of 22 regions) and department (out of 96 departments) within the French territory. We learn a latent embedding for each region and department which captures the spatial information with different levels of granularity.

3.2 Socioeconomic

Users from similar socioeconomic status tend to share similar online behavior in terms of circadian cycles. Specifically, Abitbol et al. (2018) found that people of higher socioeconomic status are active to a greater degree during the daytime and also use a more standard language. National Institute of Statistics and Economic Studies (INSEE) of France provided the population level salary for each 4 hectare square patch across the whole French territory, estimated from the 2010 tax return in France. We also use IRIS dataset provided by French government which has more coarse grained annotation for socioeconomic status. This information is mapped with the geographical coordinates of users' home location from Twitter so we can roughly ascertain the economic status of every geolocated users. We create 9 socioeconomic classes by binning the income and ensuring that the sum of income is the same for each class. We learn a latent embedding for each such class, which thus captures the variation caused by status homophily.[2]

3.3 Network

Users who are connected to each other in social networks are usually believed to share similar interests. We construct a co-mention network from the set of geolocated users as nodes and edges connecting those users who have mentioned each other at least once. We run the LINE model (Tang et al., 2015) to embed the nodes in the graph using the connectivity information and use the resulting node embedding as fixed features.

3.4 Interest

Interest feature corresponds to the set of important topics a user cares about. We obtain this information by composing a user document capturing all the words used in their posts, ranking the words in the document by the tf-idf score and selecting the top 50 of them. We then construct the user vector by summing the vectors (obtained by running word2vec on the entire corpus or geolocated tweets) corresponding to the top 50 words. We use the user vectors as fixed features.

3.5 Knowledge

Knowledge features keep track of the way the user writes and as such, it is also a summary of their content in Twitter. We learn a latent embedding for each geolocated user.

3.6 Topic

This feature associated with a tweet corresponds to the topic a tweet belongs to. Since the available corpus does not have any annotation about the topic of the tweet, we exploit the distant supervision-based idea proposed by Magdy et al. (2015) to filter geolocated tweets with an accompanying YouTube video link. We then use the YouTube public API to obtain the category of the video, which is then associated to the topic of the tweet. We learn a latent embedding for each YouTube category.

4 Proposed model

In this section we will first briefly discuss the 'Dynamic Bernoulli Embeddings' model (DBE) and then provide the details of our proposal, which uses DBE model as its backbone.

4.1 Dynamic Bernoulli Embeddings (DBE)

The DBE model is an extension of the 'Exponential Family Embeddings' model (EFE, (Rudolph et al., 2016)) for incorporating sequential changes to the data representation. Let the sequence of words from a corpus of text be represented by

[2]Some statistical pretreatments were applied to the data by INSEE before its public release to uphold current privacy laws and due to the highly sensitive nature of the disclosed data.

$(x_i, \ldots, x_N)$ from a vocabulary V. Each word $x_i \in 0, 1^V$ corresponds to a one-hot vector, having 1 in the position corresponding to the vocabulary term and 0 elsewhere. The context c_i represents the set of words surrounding a given word at position i.[3] DBE builds on Bernoulli embeddings, which provides a conditional model for each entry in the indicator vector $x_{iv} \in 0, 1$, whose conditional distribution is

$$x_{iv}|\mathbf{x}_{c_i} \sim Bern(\rho_{iv}), \tag{1}$$

where $\rho_{iv} \in (0, 1)$ is the Bernoulli probability and $\mathbf{x}_{c_i}$ is the collection of data points indexed by the context positions. Each index (i, v) in the data represents two parameter vectors, the embedding vector $\rho_v^{(t)} \in \mathbb{R}^K$ and the context vector $\alpha_v \in \mathbb{R}^K$. The natural parameter of the Bernoulli is given by,

$$\eta_{iv} = \rho_v^{\intercal}(\sum_{j \in c_i} \sum_{v'} \alpha_{v'} x_{jv'}). \tag{2}$$

Since each observation x_{iv} is associated with a time slice t_i (which is a year, in our case [4]), DBE learns a per-time-slice embedding vector $\rho_v^{(t_i)}$ for every word in the vocabulary. Thus, equation 2 becomes,

$$\eta_{iv} = \rho_v^{(t_i)\intercal}(\sum_{j \in c_i} \sum_{v'} \alpha_{v'} x_{jv'}). \tag{3}$$

DBE lets the context vectors shared across the time slices to ground the successive embedding vectors in the same semantic space. DBE assumes a Gaussian random walk as a prior on the embedding vectors to encourage smooth change in the estimates of each term's embedding,

$$\begin{aligned} \alpha_v, \rho_v^{(0)} &\sim \mathcal{N}(0, \lambda_0^{-1} I) \\ \rho_v^{(t)} &\sim \mathcal{N}(\rho_v^{(t-1)}, \lambda^{-1} I). \end{aligned} \tag{4}$$

4.2 Proposed model

In this work, we argue that the DBE model fails to accurately fit the data spanning across fewer years as it discards other explanatory variables (besides time) about the complicated processes in the language in terms of evolution and construction. These variables, which we defined in Section 3 as contextualized features, carry useful signals to understand subtle changes such as cultural drifts. Our proposed model extends DBE by utilizing these contextualized features as inductive biases.

In our setting, we represent a tweet as $t_k = (x_i, \ldots, x_N)$ belonging to user u_l. Each tuple (i, c) is associated with a set of contextualized features based on either u_l or t_k, $f_{i,m} \in \mathcal{R}^{d_m} (m = 1, \ldots, |F|)$ (where $|F|$ corresponds to the number of contextualized features). Each contextualized feature not only follows a different distribution but also has different degrees of noise (e.g., sparsity of co-mention network, geolocation inaccuracy). Hence, it is harder to unify them in a single model. We propose three ways to introduce inductive bias to the DBE model.

Unweighted sum: The simplest approach is to project all the feature embeddings to a common space and sum them up. This approach is not agnostic to the embedding vector x_i in question and consider all the contextualized features equally. Incorporating this approach, equation 3 now becomes:

$$\eta_{iv} = (\rho_v^{(t_i)} + \sum_{m=1}^{|F|} \mathbf{w}_m f_{i,m})^{\intercal} (\sum_{j \in c_i} \sum_{v'} \alpha_{v'} x_{jv'}), \tag{5}$$

where $\mathbf{w}_m$ corresponds to the learnable weights corresponding to the linear projection of $f_{i,,m}$ with size as $K \times d_m$. Note that K denotes the dimension of both context and target embedding.

Self-attention: Considering all the features equally would be wasteful for certain embedding vector x_i. Henceforth, we propose to let the network decide the important contextualized features based on self attention. This approach gives a provision to our model to handle the effect of spurious contextual signals by paying no attention. Incorporating this approach, equation 5 will now become:

$$\eta_{iv} = (\rho_v^{(t_i)} + \sum_{m=1}^{|F|} \alpha_m \mathbf{w}_m f_{i,m})^{\intercal} (\sum_{j \in c_i} \sum_{v'} \alpha_{v'} x_{jv'}), \tag{6}$$

where α_m are the scalar weights corresponding to the self-attention mechanism:

$$\alpha_m = g(f_{i,m}) = \phi(\mathbf{a}\, \mathbf{w}_m f_{i,m} + b) \tag{7}$$

where $\mathbf{a} \in \mathcal{R}^K$ and $b \in \mathcal{R}$ are learnable parameters while ϕ is a softmax.

Contextual attention: We can also make the attention mechanism to be context-dependent, that

[3] We use 2 words before and after the focal word to determine context for all our experiments.

[4] Our preliminary investigation with different time span units can be found in Appendix A.6.

DBE					Context Attn.				
2014	**2015**	**2016**	**2017**	**2018**	**2014**	**2015**	**2016**	**2017**	**2018**
jdd	*rocard*	*macron*	*macron*	*macron*	jdd	***attali***	*macron*	*macron*	*macron*
brunet	lévy	*matignon*	rugy	*élysée*	*brunet*	lévy	*matignon*	*hollande*	*élysée*
frédéric	*attali*	lejdd.fr	*hollande*	*matignon*	*dupont*	cnrs	fustige	rugy	*elysee*
elysee	*montel*	medef	*elysée*	*pétain*	*frédéric*	monarchie	renoncement	***mélenchon***	**élection**
dupont	monarchie	*élysée*	*bayrou*	interpelle	révélée	*rocard*	medef	**présidentielle**	***emmanuelmacron***

Table 1: Embedding neighborhood of 'EMMANUEL' obtained by finding closest word in each time period sorted by decreasing similarity. All named entities are italicized. Interesting words identified by the proposed model are bolded.

DBE					Context Attn.				
2014	**2015**	**2016**	**2017**	**2018**	**2014**	**2015**	**2016**	**2017**	**2018**
genesio	huitiémes	estac	*ogcnice*	*asnl*	*malcuit*	*sampaoli*	*pyeongchang*	*ogcnice*	*asnl*
génésio	*lafont*	*pyeongchang*	amical	tricolore	seri	huitiémes	*estac*	*asnl*	*eswc*
raggi	*génésio*	tgvmax	*slovaquie*	*pariez*	*tousart*	***donnarumma***	çu.e	bleuets	*carrasso*
zambo	*pyeongchang*	u20	*asnl*	affrontera	*raggi*	*lafont*	*ndombele*	*slovaquie*	tricolore
malcuit	*sampaoli*	*lrem*	bleuets	*carrasso*	*asensio*	***sertic***	*auproux*	*mennel*	**euro2016**

Table 2: Embedding neighborhood of EQUIPEDEFRANCE 'French Team' in obtained by finding the closest word in each time period sorted by decreasing similarity. All named entities are italicized. Interesting words identified by the model are bolded.

is, dependent on the embedding vector. Equation 7 then becomes:

$$\alpha_m = g(\rho_i) = \phi(\mathbf{a}_m \rho_i + b) \tag{8}$$

where $\mathbf{a}_m \in \mathcal{R}^K$ corresponds to the learnable attention parameter specific to a contextualized feature f_m.

We fit the diachronic embeddings with the *pseudo log likelihood*, the sum of log conditionals. Particularly, we regularize the pseudo log likelihood with the log priors, followed by maximization to obtain a pseudo MAP estimate. Our objective function can be summarized as,

$$\mathcal{L}(\rho, \alpha) = \mathcal{L}_{pos} + \mathcal{L}_{neg} + \mathcal{L}_{prior} \tag{9}$$

The likelihoods are given by:

$$\begin{aligned} \mathcal{L}_{pos} &= \sum_{k=1}^{|T|} \sum_{i=1}^{N} \sum_{v=1}^{V} x_{iv} log\sigma(\eta_{iv}), \\ \mathcal{L}_{neg} &= \sum_{k=1}^{|T|} \sum_{i=1}^{N} \sum_{v \in \mathcal{S}_i} log(1 - \sigma(\eta_{iv})), \end{aligned} \tag{10}$$

where $\mathcal{S}_i$ correspond to the negative samples drawn at random (Mikolov et al., 2013) and $\sigma(.)$ denote the sigmoid function, which maps natural parameters to probabilities. The prior is given by,

$$\begin{aligned} \mathcal{L}_{prior} = -\frac{\lambda_0}{2} \sum_v \|\alpha_v\|^2 - \frac{\lambda_0}{2} \sum_v \left\| \rho_v^{(0)} \right\|^2 \\ - \frac{\lambda}{2} \sum_{v,t} \left\| \rho_v^{(t)} - \rho_v^{(t-1)} \right\|^2. \end{aligned} \tag{11}$$

Language evolution is a gradual process and the random walk prior prevents successive embedding vectors $\rho_v^{(t-1)}$ and $\rho_v^{(t)}$ from drifting far apart.

The objective function established in equation 9 is learned using stochastic gradients (Robbins and Monro, 1985) with the help of Adam optimizer (Kingma and Ba, 2014). Negative samples are resampled at each gradient step. Pseudo code for training our model can be found in Appendix A.1.

5 Experiments and Results

In this section we discuss the experimental protocol, qualitative and quantitative evaluation to understand the performance of our model.

5.1 Protocol

Data: We use the French twitter dataset proposed in Abitbol et al. (2018), which is the largest collection of French tweets to date. The original dataset consists of 190M French tweets posted by 2.5M of users between June 2014 and March 2018. To be able to use socio-geographic features and assess the validity of our model, we only considered tweets from users whose home location could be identified to be in Metropolitan France. This filtering step resulted in a data set of 18M tweets from 110K users spread across 5 years. This data set was then enriched using output from the constituency-based Stanford parser in its off-the-shelf French settings (Green et al., 2011)

and from the dependency-based parser of Jawahar et al. (2018). We lowercased all the tweets, removed hashtags, mentions, URLs, emoticons and punctuations. We used 80% of the tweets from each year to train our model, split the rest equally to create validation (10%) and test set (10%). Finally, we pick the most frequent 50K words from the train set to create our vocabulary.

Baseline models: We compare our proposed model with three baseline models: **(i) Word2vec (Mikolov et al., 2013)**[5] - We use the SGNS version of Word2vec trained independently for each year with the embedding size as 100, window as 2 and the rest maintained to default; **ii) HistWords (Hamilton et al., 2016)**[6] - We use the SGNS version which is effective for datasets of different sizes and employ similar settings as the previous baseline; **(iii) DBE (Rudolph and Blei, 2018)**[7] - We use the dynamic Bernoulli embedding model (backbone of our model) with the recommended settings. We have three variants of our proposed model: no attention model (unweighted sum), self attention model and contextual attention model. Hyperparameter settings to reproduce our results can be found in Appendix A.2.

5.2 Qualitative Study

Embedding neighborhood: The goal of diachronic word embedding model is to automatically discover the changes in the usage of a word. The current usage at time t of a word w can be obtained by inspecting the nearby words of the word represented by $\rho_w^{(t)}$. From Table 1, we can observe that 'EMMANUEL' (first name of current French president) is associated with his last name ('macron') and office location ('élysée') by both DBE and proposed model. However, proposed model is able to capture interesting neighborhood by bringing words such as 'élection', 'présidentielle' and 'mélenchon' closer to 'EMMANUEL'[8]. Table 2 presents words of interest associated by our proposed model to the French football team like 'euro2016'.

Smoothness of the embedding trajectories: Since language evolution is a gradual process, the trajectory for a word tracked by a model should be changing smoothly. There are exceptions for words undergoing cultural shifts where the changes can be subtle and rapid. We plot the trajectory by computing the cosine similarity between word (e.g., MACRON) and its known, changed usage (e.g., PRESIDENT). Figure 2 shows that models relying on Bernoulli embeddings have smooth trajectories for known relations compared to other models. Despite fusing different, possibly noisy contextualized features, the trajectory tracked by our proposed model and DBE are comparably smooth.

t-SNE: Alternatively, we can overlay the embeddings from all the time slices and visualize them using dimensionality reduction technique like t-SNE (Maaten and Hinton, 2008). From Figure 3, we see a similar result where most of the words modeled by our proposed model has experienced consistent change with time.

5.3 Quantitative Study

Log Likelihood: We can evaluate models by held-out Bernoulli probability (Rudolph and Blei, 2018). Given a held-out position, a better model assigns higher probability to the observed word and lower probability to the rest. We report $\mathcal{L}_{eval} = \mathcal{L}_{pos} + \mathcal{L}_{neg}$ in Table 3. Contextual attention based model which smartly utilizes the contextualized features provides better fits to the data compared to the rest. Interestingly, the other variants of our proposed model performs poorly compared to the DBE model which suggests the importance of utilizing attention appropriately. Since all the competing methods produce Bernoulli conditional likelihoods (Equation 1), where n is the number of negative samples. We keep n to be 20 for all the methods to peform a fair comparison.

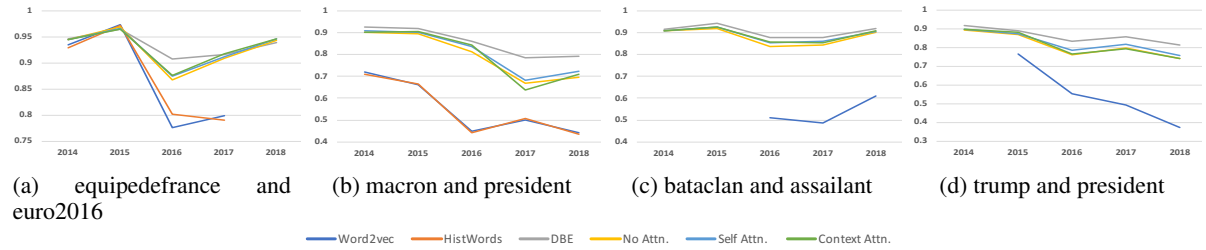
Semantic Similarity: Certain tweets are tagged with a 'category' to which it belongs (as discussed in Section 3.6). Similar to Yao et al. (2018), we create the ground truth of word category based on the identification of words in years that are exceptionally numerous in one particular category. In other words, if a word is most frequent in a category, we tag the word with that category and form our ground truth. For each category c and each word w in year t, we find the percentage of occurrences p in each category. We collect such word-time-category $\langle$w,t,c$\rangle$ triplets, avoid duplication by

[5] https://radimrehurek.com/gensim/models/word2vec.html

[6] https://nlp.stanford.edu/projects/histwords/

[7] https://github.com/mariru/dynamic_bernoulli_embeddings

[8] Emmanuel Macron became the president of France on May 2017. Jean-Luc Mélenchon stood fourth.



(a) equipedefrance and euro2016 (b) macron and president (c) bataclan and assailant (d) trump and president

Figure 2: Smoothness of word embedding trajectories vs. baseline models. High values correspond to similarity. Notice that for Word2vec model, we do not plot the results for time periods where at least one of the word of interest occurs below the minimum frequency threshold.

Model	log lik.	SS	Senti	Htag	Topic	Conv.
Word2vec	Nil	0.034	71.54	37.32	34.98	70.04
HistWords	Nil	0.042	**73.69**	36.75	36.85	70.17
DBE	-7.708	0.065	73.00	41.83	**40.01**	70.98
No Attn.	-8.059	0.058	73.22	42.11*	39.61	**71.21***
Self Attn.	-7.840	0.061	73.18	**42.19***	39.67	71.10
Context Attn.	**-7.425**	**0.068**	73.19	41.88	39.65	71.15

Table 3: Quantitative results based on log likelihood, semantic similarity and tweet classification. Higher numbers are better for all the tasks. Statistically significant differences to the best baseline for each task based on bootstrap test are marked with an asterisk. Note that we could not perform statistical significance studies for log likelihood experiment due to the large size of the test set and semantic similarity experiment due to the nature of clustering evaluation.

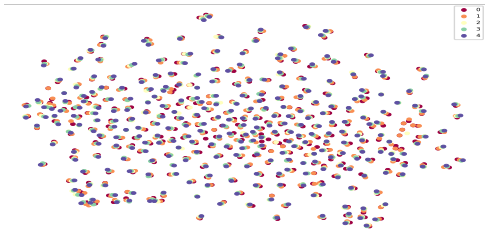
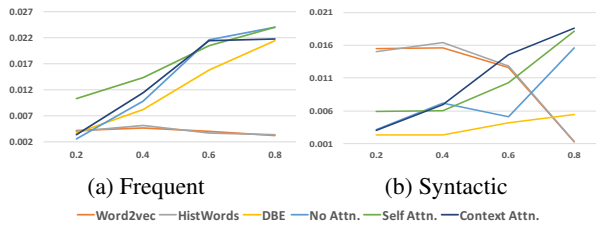


Figure 3: t-SNE visualization of mid-frequency (between 2000-2500) words for our contextual attention model.



(a) Frequent (b) Syntactic

Figure 4: Synthetic Evaluation. $p_{replacement}$ vs MRR.

keeping the year of largest strength for each w and s combination, and remove triplets where p is less than 35%. Finally, we pick top 200 words by strength from each category and create a dataset of 3036 triplets across 15 categories, where each word-year pair is essentially strongly linked to its true category. We evaluate the purity of clustering results by using Normalized Mutual Information (NMI) metric. From Table 3, we find a similar trend in the performance of our proposed model. As we see in Section 6.3, the reason our contextual attention based model excels in this task is due to its superiority in capturing semantic properties of a word.

Synthetic Linguistic Change: We can synthetically introduce the linguistic shift by introducing changes to the corpus and then evaluate if the diachronic word embedding model is able to detect those artificial drifts accurately. We follow the work done by Kulkarni et al. (2015) to duplicate our data belonging to the 2018 year 6 times (along with the extra-linguistic information), perturb the last 3 snapshots and use the diachronic embedding model to rank all the words according to their p-values. We then calculate the Mean Reciprocal Rank (MRR) for the perturbed words and expect it to be higher for models that can identify the words that have changed. To perturb the data, we sample a pair of words from the vocabulary exlcuding stop words, replace one of the word with the other with a replacement probability $p_{replacement}$ and repeat this step 100 times. We employ two types of perturbation - syntactic (where the both the words that are sampled in each step have the same most frequent part of speech tag) and frequent (where there is no restriction for the words being sampled at each step). From Figure 4, we find that DBE model is sensitive to the frequency cues from the data and fails to model subtle se-

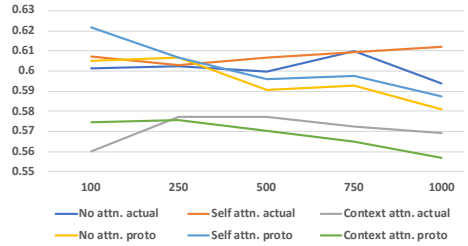


Figure 5: Change in the word's usage correlated with distance for different numbers of clusters between the 2014 and 2018 year.

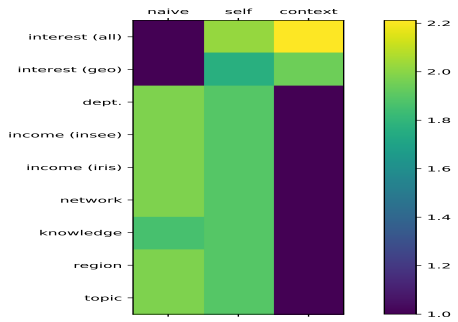


Figure 6: Importance score for each contextualized feature.

mantic shifts (e.g. for words which has evolved in its meaning without substantial change in its syntactic functionality).

Tweet Classification: We find that the existing work skips evaluating the diachronic word embeddings for a downstream NLP task. In this work we propose to test if the diachronic word embeddings can be used as features to build a temporally-aware tweet classifier.[9] We obtain a representation for a tweet by summing the embeddings for the words (belonging to the year in which tweet was posted) present in the tweet. We then train a logistic regression model and compute the F-score on the held-out instances. We establish four tweet classification tasks — Sentiment Analysis, Hashtag Prediction, Topic Categorization and Conversation Prediction (predict if a tweet will receive a reply or not) through distant supervision methods. Details of the task and dataset collection can be found in Appendix A.3. From Table 3, we find that our proposed model provides competitive performance with the baseline models for sentiment analysis and topic categorization while it outperforms them for the hashtag and conversation prediction tasks by a statistically significant margin (computed using bootstrap test (Efron and Tibshirani, 1994)). Note that there is no single best model that works for every tweet classification tasks.

6 Analysis

In this section we perform extended analysis of our proposed model to gain more insights about its functionality.

[9] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

6.1 Ablation Study

We perform ablation studies of the proposed model by considering different set of contextualized features as inductive biases, illustrated in Table 4. It is interesting to find that our model can work with a limited set of contextualized features in practice.

6.2 Law of Prototypicality

Dubossarsky et al. (2015) state that the likelihood of change in a word's meaning correlates with its position within its center. They define the prototypicality measure based on the word's distance from its cluter centroid (e.g., sword is a more prototypical exemplar than spear or dagger) and the prototypicality score reduces when the word undergoes change in its meaning. For all our models, we correlate the distance of word vector corresponding to 2014 and 2018 year with the distance the 2014 (2018) year vector moved from its cluster center. We then check if there is a positive correlation ($r > .3$). From Figure 5, we observe that there exists a positive correlation for all the variants of our model when compared to a prototypical or actual cluster centroid. Interestingly, when the cluster sizes are small (< 250), the word's meaning change is correlated with a prototypical exemplar more than a actual exemplar. On the other hand, this correlation direction gets reversed when the cluster sizes are greater than 250 and there exists more semantic areas.

6.3 Interpretation via Probing Tasks

Our tweet classification experiments (Section 5.3) demonstrated the usefulness of diachronic word embeddings as features in building a diachronic tweet classifier. Understanding the underlying properties of the tweet embeddings that enable it to outperform competing models is hard. This is why, following Conneau et al. (2018), we inves-

Task	log lik.	SS	Senti	Htag	Topic	Conv.
spatial	-7.610	0.059	73.10	42.19	39.61	71.14
income	-7.600	**0.067**	72.98	42.18	39.64	71.10
interest	-7.724	0.061	73.06	**42.21**	39.75	**71.41**
spatial & income	-7.510	0.059	73.21	42.08	39.67	71.22
spatial & interest	**-7.396**	0.059	73.11	42.14	39.77	71.23
income & interest	-7.410	0.059	73.27	42.30	39.75	71.11
spatial & income & network	-7.447	0.062	**73.35**	42.19	39.66	71.06
spatial & interest & network	-7.429	0.064	73.16	42.15	39.64	71.17
interest & income & network	-7.522	0.061	73.11	42.16	**39.82**	71.15
interest & income & network & spatial	-7.489	0.060	73.10	41.95	39.62	71.22
interest & income & network & spatial & knowledge	-7.438	0.059	73.21	41.90	39.70	71.28
interest & income & network & spatial & topic	-7.426	0.064	73.16	41.94	39.65	71.22

Table 4: Ablation Results for contextual attention model based on log likelihood, semantic similarity and tweet classification.

Model/Task (Task type)	**SentLen** (Surface)	**WC** (Surface)	**TreeDepth** (Syntactic)	**TopConst** (Syntactic)	**BShift** (Syntactic)	**Tense** (Semantic)	**SubjNum** (Semantic)	**ObjNum** (Semantic)	**SOMO** (Semantic)	**CoordInv** (Semantic)
non diachronic										
Word2vec	**84.07**	22.65	**50.34**	37.27	**50.69**	75.99	84.40	82.88	**64.40**	49.79
HistWords	83.40	**34.08**	47.51	**40.43**	49.92	**77**	**84.99**	**83.31**	64.29	**50.46**
diachronic										
DBE	73.48	46.97	43.64	31.41	**50.46**	73.34	**82.57**	82.02	64.85	50.05
No Attn.	**75.51**	46.82	**48.28**	32.78	49.15	**73.45**	82.39	82.07	65.65	49.17
Self Attn.	74.82	**47.37**	47.77	32.49	50.19	73.16	82.38	**82.18**	64.51	**50.18**
Context Attn.	75.47	46.03	47.31	**33.08**	49.98	73.05	82.10	81.81	**65.76**	49.59

Table 5: Probing task accuracies. See Conneau et al. (2018) for the details of probing tasks and classifier used.

tigate that question by setting a diagnostic classifier that probes for important linguistic features on parsed output we mentioned earlier. Those probes are based on various prediction tasks (word content, sentence length, subject or object number detection, etc.) described in (Conneau et al., 2018) and succinctly in our Appendix A.5. In 7 out of 9 tasks the use of contextual features seems to be detrimental, but the relative performance difference between our proposed models and the baseline are negligible for 5 of them. This suggests that the addition of contextualized features does not hurt the syntactic and semantic information captured by our models. Interestingly, all dynamic embeddings models are able to perform twice better in the word prediction task than a Word2vec baseline but it is unclear if those models capture language usage or actual topic prediction within a *degraded* language modeling task.

6.4 Interpretation via Erasure

Alternatively, we can directly compute the importance of a contextualized feature by observing the effects on the model of erasing (setting the weights to 0) the particular feature (Li et al., 2016). By subtracting the erased model performance on the test set from that of the original model performance and post normalization, we can establish the importance score for each feature against each version of our proposed model. Figure 6 emphasizes our finding that all contextualized features (except interest) are equally important to the performance of each variant of our proposed model.

7 Conclusion

In this work, we proposed a new family of diachronic word embeddings models that utilize various contextualized features as inductive biases to provide better fits to a social media corpus. Our wide range of quantitative and qualitative studies highlight the competitive performance of our models in detecting semantic changes over a short time range. In the future, we will consider the temporal nature of some of our contextualized features when incorporating them into our models. For example, the static social network we built can be dynamically evolving and more susceptible to accurately model underlying phenomenon.

Acknowledgments

We thank our anonymous reviewers for providing insightful comments and suggestions. This work was funded by the ANR projects ParSiTi (ANR-16-CE33-0021), SoSweet (ANR15-CE38-0011-01) and the Programme Hubert Curien Maimonide project which is part of a French-Israeli Cooperation program.

References

Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. 2018. Socioeconomic dependencies of linguistic patterns in twitter: a multivariate analysis. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1125–1134.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 380–389.

Andreas Blank and Peter Koch. 1999. *Historical semantics and cognition*. Walter de Gruyter.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2126–2136.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Proceedings of the NetWordS Final Conference on Word Knowledge and Word Usage: Representations and Processes in the Mental Lexicon, Pisa, Italy, March 30 - April 1, 2015.*, pages 66–70.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21. Association for Computational Linguistics.

Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. 2019. Incorporating demographic embeddings into language understanding. *Cognitive Science*, 43(1).

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision.

Spence Green, Marie-Catherine De Marneffe, John Bauer, and Christopher D Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with french. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 725–735. Association for Computational Linguistics.

Joachim Grzega and Marion Schoener. 2007. English and general historical lexicology.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Ganesh Jawahar, Benjamin Muller, Amal Fethi, Louis Martin, Eric Villemonte de la Clergerie, Benoît Sagot, and Djamé Seddah. 2018. ELMoLex: Connecting ELMo and lexicon features for dependency parsing. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 223–237, Brussels, Belgium. Association for Computational Linguistics.

Dan Jurafsky. 2018. *Speech & language processing*, 3rd edition. Currently in draft.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the Workshop on Language Technologies and Computational Social Science@ACL 2014, Baltimore, MD, USA, June 26, 2014*, pages 61–65.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 625–635.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1384–1397.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Walid Magdy, Hassan Sajjad, Tarek El-Ganainy, and Fabrizio Sebastiani. 2015. Distant supervision for tweet classification using youtube labels. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 638–641.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Herbert Robbins and Sutton Monro. 1985. A stochastic approximation method. In *Herbert Robbins Selected Papers*, pages 102–109. Springer.

Maja R. Rudolph and David M. Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1003–1011.

Maja R. Rudolph, Francisco J. R. Ruiz, Stephan Mandt, and David M. Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 478–486.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, abs/1811.06278.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 1067–1077.

Jason Weston, Sumit Chopra, and Keith Adams. 2014. #tagspace: Semantic embeddings from hashtags. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1822–1827.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 673–681.

A Appendices

A.1 Pseudo code for the training algorithm

Pseudo code for training our proposed model is presented in Algorithm 1.

Algorithm 1 : Training algorithm for the proposed diachronic word embedding model

Input: Tweets X^t of size m_t from T time slices, contextual features f_m, context size c, embedding size K, number of negative samples n, number of minibatch fractions m, initial learning rate η, precision λ, vocabulary size V, smoothed unigram distribution $\hat{\rho}$.

for v from 1 **to** V **do**
  Initialize α_v and $\rho_v^{(T)}$ with $\mathcal{N}(0, 0.01)$
end for
for m from 1 **to** $|F|$ **do**
  if f_m is learnable **then**
    Initialize f_m with $\mathcal{U}(0, 1)$
  end if
end for
for number of passes over the data **do**
  for number of minibatch fractions m **do**
    for t from 1 **to** T **do**
      for i from 1 **to** $\frac{m_t}{m}$ **do**
        Sample $c+1$ consecutive words from a random tweet $X^{(t)}$ and construct: $C_i^{(t)} = \sum_{j \in c_i} \sum_{v'} \alpha_{v'} x_{jv'}$
        Compute contextualized features: $F_i^{(t)} = \sum_{m=1}^{|F|} \alpha_m \mathbf{w}_m f_{i,m}$ Draw a set $S_i^{(t)}$ of n negative samples from $\hat{\rho}$.
      end for
    end for
    Update the parameters $\theta = \alpha, \rho, f_m, w, a, b$ by ascending the stochastic gradient
    $\nabla_\theta \{ \sum_{t=1}^{T} m \sum_{i=1}^{\frac{m_t}{m}} (\sum_{v=1}^{V} x_{iv}^{(t)} log\sigma((\rho_v^{(t)} + F_i^{(t)})^T C_i^{(t)})$
    $+ \sum_{x_j \in S_i^{(t)}} \sum_{v=1}^{V} log(1 - \sigma((\rho_v^{(t)} + F_i^{(t)})^T C_i^{(t)})$
    $- \frac{\lambda_0}{2} \sum_v \|\alpha_v\|^2 - \frac{\lambda_0}{2} \sum_v \left\|\rho_v^{(0)}\right\|^2$
    $- \frac{\lambda}{2} \sum_{v,t} \left\|\rho_v^{(t)} - \rho_v^{(t-1)}\right\|^2 \}.$
  end for
end for
We utilize Adam (Kingma and Ba, 2014) to set rate η.

A.2 Hyperparameter settings

We follow the hyperparameter search space provided by Rudolph and Blei (2018) to find the best configuration of our model. Before training our model, we initialize the parameters with one epoch fit of non-diachronic Bernoulli embedding model (as defined in Equation 2 in the paper). We then train our model for 9 more epochs. We fix the embedding dimension to 100, context size to 2 and number of negative samples to 20. We select the initial learning rate $\eta \in [0.01, 0.1, 1, 10]$, minibatch size $m \in [0.001N, 0.0001N, 0.00001N]$ (where N is the number of training records), the precision on context vectors and initial dynamic embeddings $\lambda \in [1, 10]$ ($\lambda_0 = \lambda/1000$). We use the conditional likelihood metric (as discussed in Section 5.3) to sweep over the search space and select the best hyperparameters.

A.3 Tweet Classification Details

We will list down the details of tweet classification tasks where the data comes from our corpus.

- **Sentiment Analysis** - This is a binary task to classify the sentiment of the tweet. Following Go et al. (2009), we create a balanced dataset by tagging a tweet as positive (negative) if it contains only positive (negative) emoticons. We remove the emoticons from the tweets to avoid bias.
- **Hashtag Prediction** - This multiclass classification task is to identify the hashtag present in the tweet. Following Weston et al. (2014), we identify the most frequent 100 hashtags from the corpus, keep the tweets that contain exactly one occurrence of the frequent hashtag, remove the hashtag from the tweet and predict them.
- **Topic Categorization** - This multiclass classification task is to identify the topical category to which a tweet belongs to. Following Magdy et al. (2015), we filter the tweets that has a YouTube video associated with it, query the video category using the public YouTube API and associate that to the topical category of the tweet.
- **Conversation Prediction** - This binary task is to classify if a tweet will receive a reply or not. Following Elazar and Goldberg (2018), we tag the tweet as a conversational tweet if it has at least a mention ('@') in it, otherwise it's a non-conversational tweet. We remove the mentions from the tweets to avoid bias.

A.4 Ablation Results

We perform ablation studies of the no attention and self attention variant of the proposed model by considering different set of contextualized features as inductive biases, illustrated in Table 6.

A.5 Probing Task Description

In this section we will describe briefly the set of probing tasks (proposed in Conneau et al. (2018)) used in our study.

- **SentLen** - The goal for the classification task is to predict the tweet length which has been binned in 6 categories with lengths ranging in the following intervals: $(5-8), (9-12), (13-16), (17-20), (21-25), (26-28)$.
- **WC** - This classification task is about predicting which of the target words appear on the given tweet.
- **TreeDepth** - In this classification task the goal is to predict the maximum depth of the tweet's syntactic tree (with values ranging from 5 to 12).
- **TopConst** - The goal of this classification task is to predict the sequence of top constituents immediately below the sentence (S) node. The classes are given by the 19 most common top-constituent sequences in the corpus, plus a 20th category for all other structures.
- **BShift** - In this binary classification task the goal is to predict whether two consecutive tokens within the tweet have been inverted or not.
- **Tense** - The goal of this task is to identify the tense of the main verb of the tweet.
- **SubjNum** - The goal of this task is to identify the number of the subject of the main clause.
- **ObjNum** - The goal of this task is to identify the number of the subject on the direct object of the main clause.
- **SOMO** - This task classifies whether a tweet occurs as-is in the source corpus, or whether a randomly picked noun or verb was replaced with another form with the same part of speech.
- **CoordInv** - This task distinguishes between original tweet and tweet where the order of two coordinated clausal conjoints has been inverted purposely.

A.6 Selection of time span unit

We performed preliminary experiments with DBE model to identify the time span unit that best fits the data. As shown in Table 7, DBE model fits the data well in terms of log likelihood metric when the time span unit is year.

Time span unit	Yearly	Monthly	Quarterly	Half-yearly
Log lik.	**-5.7323**	-7.1055	-6.4004	-6.0768

Table 7: Log likelihood scores of DBE model with varying time span units.

Task	log lik.	SS	Senti	Htag	Topic	Conv.
No Attention						
spatial	-7.8481	0.0583	73.11	42.1	39.76	71.16
income	-7.8407	0.0616	73.17	41.99	**39.80**	71.22
interest	-7.9704	0.0596	73.24	42.11	39.72	71.17
spatial & income	-7.8407	**0.0718**	73.18	42.07	39.67	71.17
spatial & interest	-7.9774	0.0581	73.27	42.05	39.69	71.14
income & interest	-7.9601	0.0620	**73.3**	42.08	39.68	71.14
spatial & income & network	**-7.7735**	0.0614	73.2	42.13	39.78	71.17
spatial & interest & network	-8.0061	0.0613	73.27	42.17	39.61	71.14
interest & income & network	-8.0170	0.0605	73.22	42.1	39.71	71.17
interest & income & network & spatial	-8.0561	0.0587	73.29	42.18	39.67	71.15
interest & income & network & spatial & knowledge	-8.0734	0.0620	**73.3**	**42.19**	39.7	**71.24**
interest & income & network & spatial & topic	-8.0739	0.0639	73.28	42.13	39.62	71.15
Self Attention						
spatial	-7.8260	0.0624	73.11	41.95	39.75	71.09
income	-7.8248	0.0577	73.13	41.98	39.78	**71.17**
interest	-7.7986	0.0602	73.21	42.14	**39.83**	71.13
spatial & income	-7.8383	**0.0641**	73.08	42.02	**39.83**	71.14
spatial & interest	-7.7874	0.0625	73.2	42.11	39.71	71.12
income & interest	-7.7796	0.0635	73.2	**42.18**	39.75	71.11
spatial & income & network	-7.8613	0.0609	73.09	42.07	39.77	71.19
spatial & interest & network	**-7.7558**	0.0611	73.13	42.1	39.68	71.07
interest & income & network	-7.8432	0.0607	73.11	42.13	39.48	71.09
interest & income & network & spatial	-7.8414	0.0609	73.15	42.16	39.58	71.04
interest & income & network & spatial & knowledge	-7.8554	0.0618	73.2	42.15	39.58	71.06
interest & income & network & spatial & topic	-7.8208	0.0575	**73.22**	42.13	39.58	71.07

Table 6: Ablation results based on log likelihood, semantic similarity and tweet classification.