Biographically Relevant Tweets – A New Dataset, Linguistic Analysis and Classification Experiments

Michael Wiegand
Digital Age Research Center (D!ARC)
Alpen-Adria-Universität Klagenfurt
AT-9020 Klagenfurt, Austria
michael.wiegand@aau.at

Rebecca Wilm
Institute of Computational Linguistics
Heidelberg University
D-69120 Heidelberg, Germany
wilm@cl.uni-heidelberg.de

Katja Markert
Institute of Computational Linguistics
Heidelberg University
D-69120 Heidelberg, Germany
markert@cl.uni-heidelberg.de

Abstract

We present a new dataset comprising tweets for the novel task of detecting biographically relevant utterances. Biographically relevant utterances are all those utterances that reveal some persistent and non-trivial information about the author of a tweet, e.g. habits, (dis)likes, family status, physical appearance, employment information, health issues etc. Unlike previous research we do not restrict biographical relevance to a small fixed set of pre-defined relations. Next to classification experiments employing state-of-the-art classifiers to establish strong baselines for future work, we carry out a linguistic analysis that compares the predictiveness of various high-level features. We also show that the task is different from established tasks, such as aspectual classification or sentiment analysis.

1 Introduction

Since its beginning, the web has been a valuable data source for natural language processing (NLP) applications. Particularly social media contain data complementary to what previous (news) text corpora could offer (Farzindar and Inkpen, 2015), such as information on ordinary people.

This work focuses on biographical information on ordinary users of Twitter. The **task** is to identify tweets that contain biographically relevant information on the author of a tweet. We frame this as a **binary text classification task**: We distinguish between *biographically relevant tweets* (1)-(5) and *biographically irrelevant tweets* (6)-(10).

(1) *relevant:* I'm a grown man in my mid-30s with no children
(2) *relevant:* Today was the first day at Deckers my manager told me good job lmao.
(3) *relevant:* one of the reasons i regularly wear @AmericanOutlaws kit is because it makes a political statement
(4) *relevant:* i'm actually rather shy in real life
(5) *relevant:* Brenda Fricker will always be my favourite Irish actress
(6) *irrelevant:* It's after 1 a.m. here and I am stupidly awake
(7) *irrelevant:* @USER i wanna buy your mum a beverage.
(8) *irrelevant:* I heard they're changing a dollar for corn at the Tampa Bay stadium, that's a buck an ear.
(9) *irrelevant:* I'm glad I stayed home last night but I'm sad I was alone lmfaooooooooo
(10) *irrelevant:* Headache? I've had a headache all day xx

Unlike previous work (Sekine and Artiles, 2009; Ji et al., 2010; Garcia and Gamallo, 2015), we **do not use a pre-specified list of biographical relations**. Instead, **we consider everything as a biographically relevant relation if it reveals some persistent and non-trivial information on the user authoring a particular tweet**. This may include simple information on age and family status (1), (major) life events (2), habits (3), personality traits (4), (dis)likes (5), etc. Tweets may offer a plethora of biographical information about their users and we believe that such unrestricted setting is faithful to the nature of the data that we consider.

Biographical information on ordinary users would be valuable for various applications. Providers of social-media sites could support their users with effectively building up their social network, for example, by recommending friends based on shared interests (Xie, 2010). Another obvious application with a high commercial potential is product recommendation (Ricci et al., 2011). Personalized advertisement using biographical information about a user (e.g. recommending baby care

Proceedings of the 29th International Conference on Computational Linguistics, pages 3669–3679
October 12–17, 2022.

products for parents-to-be) may be much more effective than undirected general advertisement.

Biographical information may also be used for applications that are **questionable from an ethical point of view**. For instance, potential employers could do background checks on job seekers. Similarly, insurance companies could screen potential insurants before offering them a particular insurance. Specific biographic information could well have a negative impact on the customers. For example, a health insurance might decline a customer an offer if they knew about their unhealthy life style.

We are fully aware of these ethical issues and we want to make clear that our aim is not to provide basic research for technologies that could be used for such questionable purposes to evolve. On the contrary, we could **envisage scenarios that would help users to protect the above misuse of their personal data** (Malandrino et al., 2013). For instance, the functionality to identify biographically relevant information in user-generated content may be incorporated in an application that automatically anonymizes such sensitive text passages (Medlock, 2006; Lison et al., 2021) or, at least, warns users prior to posting such content that they are about to disclose sensitive material.

The question may arise **why we aim for a binary classification task and do not specify the subtype of biographical information**. We consider our task just as the first step in the pipeline of processing biographically related material. Obviously, processing all microposts of a social-media site manually would be a prohibitive task. A great reduction of work could be achieved by selecting all biographically-relevant material automatically, as proposed in this paper. Only a small proportion of tweets within a user timeline from Twitter – our estimate based on our data is that this is about 33% – actually contains such information.

The aim of this research is two-fold. First, we want to investigate whether biographically relevant utterances can reliably be identified. Second, we also want to determine what types of biographically relevant information one may encounter on Twitter.

Our **contributions** are the following:

- We introduce the novel task of detecting biographically relevant utterances.
- We present a new dataset for this task.
- We report on classification performance using state-of-the-art supervised classifiers establishing strong baselines for future work.
- By testing classifiers trained on related tasks on our task, we can show that our novel task is different from previous tasks.
- We carry out a linguistic analysis on our new dataset which sheds more light on the nature of biographically relevant utterances.

All data created as part of this research are made available for research purposes upon request.[1]

2 Related Work

A substantial body of previous research on the extraction of biographical information using NLP focuses on pre-defined relations, such as *hasBirthDate(<person>,<date>)* (Sekine and Artiles, 2009; Ji et al., 2010; Garcia and Gamallo, 2015). In this paper, we depart from this setting. Our observation is that in social media there is plenty of biographically relevant information that does not easily fit into this finite set of relations:

(11) my parents were complete idiots and I try to do the opposite of anything they would ever do

As a data source for extracting biographical information, Wikipedia[2] has, so far, been the most frequently used corpus (Garera and Yarowsky, 2009; Garcia and Gamallo, 2015; Plum et al., 2019). Consequently, previous research typically focused on celebrities or people who are notable in some way since it is this group of people who are predominantly represented in Wikipedia. By focusing on social media (i.e. Twitter), we are able to extract biographical information from *ordinary people*. This also means that we are able to widen the scope of biographically relevant information.

The task addressed in this paper also bears some relation to multi-document (biography) summarization (Zhou et al., 2004; Biadsy et al., 2008) which is typically framed as a two-step approach: In the first step, biographically relevant utterances (from different documents) are identified while in the second step, a summary of these extracted utterances is generated by eliminating redundancies and creating a plausible order. Our task matches the first step although in the context of summarization previous work only dealt with pre-defined relations.

[1] Due to the sensitivity of the data, access is only granted if an outline of the intended research is submitted via email to the authors and that outline is considered ethically sound. The annotation guidelines we used for producing our dataset, however, can be downloaded directly: https://github.com/miwieg/biographical_relevance_guidelines

[2] https://en.wikipedia.org

number of timelines	27
number of tweets	14,483
average number of tweets per timeline	536.4
average number of tokens per tweet	17.7
type-token ratio	8.7%
category: biographically relevant tweets	4,724 (32.6%)
category: biographically irrelevant tweets	9,759 (67.4%)

Table 1: Statistics of the BRT-dataset.

With regard to social media, there have been related research efforts in the area of *author profiling* (Rao et al., 2010; Zamal et al., 2012; Rangel et al., 2013). These works focus on demographic categories, such as predicting gender, age or personality traits. Moreover, there has been research on extracting *major life events*. Such events are environmental circumstances that have an identifiable onset and ending and may carry the potential for altering an individual's present state of mental or physical well-being (Goodyer, 2001). Examples are getting married or finding a new job (Li et al., 2014; Li and Cardie, 2014; Dickinson et al., 2015). As we will provide evidence later, major life events and demographic categories only form a subset of the relations we consider as biographical relevance.

In their recent study, Saha et al. (2021) examine what life events are disclosed on social media. The authors establish that a significant number of such events can be found. Unlike our work in which we address both data analysis and classification experiments, Saha et al. (2021) present a purely descriptive study. Moreover, that work is carried out on data collected from Facebook, while our work is carried out on data crawled from Twitter. While all life events represent biographically relevant information, the set of biographically relevant utterances considered in this paper also includes relations beyond life events, e.g. likes or habits.

In the area of privacy protection (Malandrino et al., 2013), there has also been research using NLP. Mao et al. (2011) present a descriptive study examining properties of three types of leaks on Twitter: divulging vacation plans, tweeting under the influence of alcohol and revealing medical conditions. Cappellari et al. (2017) present a tool for privacy detection trained on 500 tweets manually labeled as *private* or *not-private* using SVM. Neerbek et al. (2017) introduce a system that distinguishes between *sensitive* vs. *non-sensitive* information based on recursive neural networks (Irsoy and Cardie, 2017). Information about the training data used are not provided. Canfora et al. (2018) establish a set of heuristic rules operating on dependency parses for the detection of privacy leaks. The latter are approximated by mentions of locations or emotions. The approach is examined on a random set of 856 Facebook statuses. Our work differs from these works in that our focus is not just on sensitive data but all different types of biographical information. For example, habits, personality traits or likes and dislikes are biographical information but they need not be categorized as (highly) sensitive data. Moreover, our dataset with about 14k tweets that has been annotated manually is considerably larger. It has also been constructed in a more principled and less biased way. For example, Mao et al. (2011) create their dataset by sampling for specific topic keywords while our raw data represent timelines of users (§3). Finally, our classification experiments are carried out using state-of-the-art classifiers which are much more robust than those in previous work (e.g. SVM). Unfortunately, none of those previous datasets on privacy protection are publicly available, so we could not consider them in our evaluation.

3 Data

We now describe the creation of our novel dataset to detect **b**iographically **r**elevant **t**weets, henceforth referred to as **BRT-dataset**. We did not sample our dataset using specific keywords since this results in topic biases (Wiegand et al., 2019). Neither did we randomly sample tweets since, as our exploratory experiments revealed, the tweets would have a bias towards topics predominant during collection time, such as COVID-19. Instead, we extracted timelines of users. Using the Twitter API[3], we selected users from Australia, Canada, Ireland, New Zealand, the UK and the USA who had recently posted a tweet containing the first-person singular pronoun. Since we could only annotate a small number of timelines, we tried to come up with a subset of very different users (having different ages, genders, ethnicities, sexual orientations, political beliefs etc.). Thus, we hope to appropriately account for the demographic diversity of our society.

In order to reduce the manual annotation, we got rid of trivial cases of irrelevant tweets, i.e. tweets that do not contain any mention of the author of the tweet itself. That is, we excluded tweets without a mention of the first-person pronoun.[4] Since we

[3] https://developer.twitter.com/en/docs/twitter-api

[4] On a set of about 2,000 tweets that were thus discarded,

did not want to discard cases of *implicit protagonists* (12) (Regneri et al., 2011), which despite the absence of such pronoun refer to the author, we applied some further heuristics (e.g. by including declarative sentences that begin with a verb).

(12) taking gym selfies#joined a gym today!

Each tweet of those timelines was manually labeled. For practical reasons, tweets were annotated in isolation. As already discussed in §1, we define a tweet as biographically relevant if it reveals some persistent and non-trivial information on the user authoring a particular tweet. Tweets were to be labeled as biographically relevant no matter whether the information was mentioned explicitly (13) or implicitly (14). In (13) the author explicitly states their pride to be Americans. However, in (14), we infer the author's reading enthusiasm since (s)he asks for book recommendations. In BRT, 28% of the biographically relevant utterances are implicit.

We insisted that the biographically relevant information should be unambiguous. For instance, (15) could imply that the author is a regular churchgoer in which case the tweet would be biographically relevant. However, since there are also several other possible explanations (for instance, the author might just be attending a wedding) we did not label such cases as biographically relevant.

(13) my dad has always hung an American flag outside our house, cause we are PROUD TO BE AMERICANS
(14) uhm any book recommendations? i see a day trip to barnes n noble in my near future for mental health care
(15) #Omw to church. :)

On a random subset of 300 tweets we measured the agreement between our annotator (a member of the department of one of the co-authors) and one co-author of this paper. We obtained a substantial inter-annotator agreement of Cohen's $\kappa = 0.73$ (Landis and Koch, 1977).

Table 1 provides some summarizing statistics of our final dataset. The dataset consists of more than 14,000 tweets. Only about one third of them are considered biographically relevant.

4 High-Level Features

In this paper, we follow a supervised learning approach. Among the different classifiers we consider, we also implement a feature-based classifier using high-level features. These features are also used for a descriptive analysis on our BRT-dataset. In the following, we describe the feature set we devised.

4.1 Aspectual Classification (ASPECT)

By aspect, we understand how an action, state or event denoted by a verb extends over time.[5] We consider the aspectual categories as proposed by Friedrich and Pinkal (2015) who distinguish between static aspect, i.e. clauses expressing states (16), episodic aspect, i.e. clauses expressing information about events (17), and habitual aspect, i.e. clauses expressing regularities (18).

(16) I like coffee. (*invented example*)
(17) I bought some coffee right now. (*invented example*)
(18) I usually drink coffee after lunch. (*invented example*)

While biographically relevant utterances can co-occur with all 3 aspectual categories, we assume that they are not equally distributed across these categories. For example, most events (=episodic aspect) should not be life-changing ones and therefore be biographically irrelevant. On the other hand, habitual aspect may often refer to habits or hobbies that tell us a lot about a person and therefore could qualify as biographical relevance.

As a tool to determine the aspectual category of an utterance, we use *sitent* (Friedrich et al., 2016).

4.2 Sentiment Analysis (SENTI)

Intuitively, there seems to be a relation between sentiment (Liu, 2012) and biographical relevance as people's likes and dislikes are typically expressed by positive and negative sentiment (19). However, not all instances of positive and negative sentiment may be biographically relevant. For instance, (dis)likes that are shared by everyone (20) are considered irrelevant since they represent trivial information. We run *TweetEval* (Barbieri et al., 2020) on the tweets of BRT. In addition to a feature that indicates the predicted sentiment category (i.e. positive, negative or neutral), we included features that reflect the range in which the confidence probability score of the prediction for a particular tweet falls. We divide that score into bins of size 0.1.

(19) I honestly hate group assignments.
(20) I really hate doing something that I don't want to do

we only identified 4% biographically relevant instances. Therefore, we can conclude that our filtering heuristic only misses a negligible proportion of actual biographically relevant tweets.

[5] https://en.wikipedia.org/wiki/Grammatical_aspect

4.3 Emotion Classification (EMOTION)

We also want to investigate in how far emotion categories, such as *joy*, *fear* or *surprise*, correlate with biographical relevance. Intuitively, certain emotional states may coincide with the author exhibiting some mental condition (21). We consider the *NRC emotion lexicon* (Mohammad and Turney, 2013) and associate each tweet with the set of emotion categories that are triggered by the respective emotion words contained in the tweet.

(21) seasonal desperation and regular depression ready to fuck me over

4.4 Supersenses (SUPER)

WordNet (Miller et al., 1990) groups each synset into one of 45 supersenses[6] which represent coarse-grained semantic categories, such as *noun.food* or *verb.motion*. For a linguistic analysis, supersenses can be informative as they may indicate which semantic concepts correlate with biographical relevance. We represent each tweet by the supersenses associated with the words that occur in it.[7]

4.5 Part-of-Speech Information (POS)

We investigate whether biographical relevance displays a specific distribution of part-of-speech (POS) tags. Each tweet is associated with the set of POS tags of the words occurring in it. We use the POS tagger by Owoputi et al. (2012), which has been optimized for Twitter.

4.6 Family-Member Wordlist (FAMILY)

A frequently occurring subtype of biographical relevance are family relations, e.g. does the author have children, are they married, do they have any siblings etc. Since family members and partners represent a clear-cut concept, we compiled a list of 105 lemmas expressing these relations. We implemented a look-up feature that indicates whether any of these lemmas could be found in a tweet.

4.7 Meta Features (META)

We also want to determine the effectiveness of meta information for our task. We distinguish between 3 types of meta features described below. All features are binary features. This design decision was made since it substantially facilitates our linguistic analysis in §5.[8] Table 2 summarizes all meta features we examine in this work. Several of these features represent rankings. They were discretized by dividing the range of values into bins. Table 2 also indicates how the discretization is conducted.

Tweet-Level Meta Features. These are features that refer to an individual tweet. Two of them need to be explained in more detail since they required a special form of normalization. It concerns the feature that counts the number of likes assigned to a tweet (*META_TWEET_is_among_top_n_likes*) and the feature that counts the number of the times a feature has been retweeted (*META_TWEET_is_among_top_n_retweets*). Since different users typically receive quite a different number of likes (or number of retweets), we normalized that number by the average number of likes (or retweets) a particular user obtained. The resulting ranking therefore reflects whether a tweet received more or fewer likes (or retweets) than expected.

Thread-Level Meta Features. These are features referring to the thread a tweet is situated in.

User-Level Meta Features. These are features referring to the user that authored a particular tweet.

5 Linguistic Analysis

Table 3 shows for both classes the 20 high-level features (§4) with the highest precision. The strongest feature to correlate with biographical relevance is the word list referring to family relations. This result is quite intuitive. However, since that feature only fires in 863 of the more than 4,000 cases of biographical relevance on BRT, we can conclude that our data also contain many other forms of biographical relevance. A considerable proportion of further predictive features concern supersenses. The most predictive one concerns pertainyms (*adj.pert*). These seem to be relational adjectives that often relate to demographic information (e.g. *my economic situation*, *British nationality*). Further nouns expressing feelings (*noun.feeling*) may refer to likes, interests or mental health (e.g. *love*, *anxiety*). Temporal nouns (*noun.time*) may refer to *birthday*, *age* or *anniversary* etc. In terms of POS-tags, the possessive pronoun is the strongest. It is often part of frequently occurring biographically relevant

[6] https://wordnet.princeton.edu/documentation/lexnames5wn

[7] Due to the lack of robust word-sense disambiguation, we represent each word as the union of synsets containing it.

[8] It is much more straightforward to rank binary features according to the predictiveness towards a particular class than a feature set containing non-binary (continuous) features.

Abbreviation	Explanation	Info. about Discretization
META_TWEET_is_answer	is tweet an answer	
META_TWEET_is_retweet	is tweet a retweet	
META_TWEET_is_quote	is tweet a quote tweet	
META_TWEET_is_posted_on_weekend	has tweet been posted during weekend	
META_TWEET_is_among_top_n_likes	is tweet among top n tweets with the most likes	n: [200, 500, 1000, 2000]
META_TWEET_is_among_top_n_retweets	is tweet among top n tweets with the most retweets	n: [200, 500, 1000, 2000]
META_THREAD_is_thread_length_n	does thread have n tweets	n: [0, 1, 2, 3, 4, 5 or more]
META_THREAD_has_n_contributors	does user thread have n contributors	n: [0, 1, 2, 3, 4, 5 or more]
META_THREAD_has_n_tweets_from_user	did user post n tweets in this thread	n: [0, 1, 2, 3, 4, 5 or more]
META_USER_is_country_code_x	is country code of user x (e.g. UK, US, CA etc.)	
META_USER_has_n_to_m_likes	does user have between n and m likes	bins (n, m) are log. scaled
META_USER_has_n_to_m_followers	does user have between n and m followers	bins (n, m) are log. scaled
META_USER_is_following_n_to_m_users	is user following between n and m users	bins (n, m) are log. scaled
META_USER_posted_n_to_m_tweets	did user post between n and m tweets	bins (n, m) are log. scaled
META_USER_included_in_n_to_m_lists	is user included in n to m lists	bins (n, m) are log. scaled
META_USER_has_default_profile	does tweet have user default profile	

Table 2: Overview of meta features.

phrases, such as *my wife*, *my dog* or *my car*. Finally, we also find a few predictive meta features: If a tweet of a user is more often *liked* or *retweeted* than the average tweet (of the user), then it is likely to be biographically relevant.

With regard to biographical irrelevance, we find many fewer linguistic features and predominantly meta features. The most predictive feature is whether a tweet is a retweet. Given that tweets are typically retweeted by users other than the author of the original tweet[9], there are (at least) three possible reasons for the predictiveness of this feature. First, it is possible that users do not disclose or forward biographical information of other people often. Second, maybe people are more interested in writing about themselves. Third, as we only include tweets written in the first person, we might miss biographical information in the second or third person.

Most of the other features predictive for biographical irrelevance are meta features that refer to the thread of the tweet. These features are also correlated to each other. If a thread is lengthy or has many users contributing to it, then this means that there is some deeper discussion about it. Obviously biographically relevant content is less likely to be the centre of such discussions. Apart from the meta features, sentiment information is the most predictive linguistic feature for biographical irrelevance. Tweets that have been classified as neutral with a high confidence are likely to be biographically irrelevant. This is plausible since likes and dislikes (which are mostly biographically relevant) are positive or negative.

Neither aspectual categories (§4.1) nor emotion categories (§4.3) occur in either of the rankings. The latter category was observed frequently in both biographically relevant tweets (22) and biographically irrelevant tweets (23).

(22) seasonal desperation and regular depression ready to fuck me over

(23) @SmartRachael It's a good day. I'm happy. X

As far as aspectual classification is concerned, we noted heavy noise in the output of *sitent*, i.e. the tool that provided us with an aspectual analysis of our data.

6 Classification Experiments

We now report on our classification experiments.

Baselines. As baselines, we consider a **majority-class classifier** that always predicts biographical irrelevance. Furthermore, we consider a **sentiment classifier** which predicts biographical relevance if a tweet either conveys positive or negative rather than neutral sentiment. (As in §4.2, we obtain sentiment information using *TweetEval*.) Our third baseline is an **aspectual classifier** which predicts biographical relevance if a tweet conveys *habitual* aspect. (As in §4.1, we obtain the aspectual classification using *sitent*.) Our fourth baseline is a **classifier using meta-information**. Motivated by our analysis from §5, it predicts biographical relevance if a tweet is neither a retweet, a quote tweet nor represents the answer to a tweet. The more predictive this baseline is, the less important we should consider NLP-based information for this task.

[9]Please note that there is a difference between the feature *META_TWEET_is_among_top_200_retweets* which is predictive for biographical relevance and *META_TWEET_is_retweet* which is predictive for biographical irrelevance. The latter feature indicates that a tweet is a retweet itself while the former is related to the number of times the tweet has been retweeted. Apparently, a retweet per se is an indication of biographical irrelevance while if a tweet of a particular user has been retweeted disproportionally often, then there is a tendency of the tweet being biographically relevant.

Biographically Relevant			Biographically Irrelevant		
Feature	**Prec**	**Freq**	**Feature**	**Prec**	**Freq**
FAMILY	73.8	863	META_TWEET_is_retweet	83.3	216
SUPER_adj.pert	53.1	682	META_THREAD_is_thread_length_4	80.8	318
META_USER_has_5000_to_9999_followers	52.3	285	META_THREAD_has_4_tweets_from_user	80.7	57
SUPER_noun.phenomenon	50.6	1,557	SENTI_neutral_confidence_range_0.9_to_1.0	75.7	66
POS_PRP$	48.4	4,672	META_TWEET_is_quote	75.2	1,587
META_TWEET_is_among_top_500_likes	47.8	500	META_THREAD_has_2_tweets_from_user	75.1	438
POS_RRB	47.5	383	META_THREAD_has_2_contributors	74.7	2,251
SUPER_noun.Tops	47.4	1,855	META_THREAD_is_thread_length_3	74.6	574
SUPER_noun.feeling	47.2	1,375	META_USER_has_1,000_to_1,999_likes	74.5	2,139
META_USER_posted_20000_to_49999_tweets	46.7	197	META_TWEET_is_answer	73.8	6,325
META_TWEET_is_among_top_200_likes	46.5	200	META_THREAD_is_thread_length_2	73.8	1,581
META_USER_is_following_1000_to_1999_users	46.0	491	META_THREAD_has_1_tweet_from_user	73.7	2,091
META_TWEET_is_among_top_200_retweets	46.0	200	META_USER_posted_1000_to_1999_tweets	73.5	898
SUPER_noun.relation	46.0	774	META_THREAD_is_thread_length_1	73.5	3,248
POS_LRB	45.8	393	META_THREAD_has_1_contributor	73.4	3,581
SUPER_noun.time	45.8	4,306	META_THREAD_has_4_contributors	73.1	130
POS_HT	45.5	1,065	POS_USR	73.0	7,061
POS_RBR	45.4	359	META_USER_is_following_0_to_99_users	72.7	2,532
SUPER_noun.group	45.2	1,629	META_USER_is_country_code_US	71.6	6,365
POS_JJS	44.8	411	META_USER_is_following_200_to_499_users	71.6	2,450
random precision	32.6		**random precision**	67.4	

Table 3: Top 20 features with highest precision for the different classes.

		Randomized Folds				Unseen Timelines			
Classifier	**Supervised Classifier?**	**Acc**	**Prec**	**Rec**	**F1** (*std*)	**Acc**	**Prec**	**Rec**	**F1** (*std*)
majority-class baseline	no	67.4	33.7	50.0	40.3	67.4	33.7	50.0	40.3
sentiment baseline	no	45.0	50.5	50.5	50.5	45.0	50.4	50.5	50.4
aspectual baseline	no	65.3	55.7	53.1	54.3	65.3	55.7	53.1	54.4
meta baseline	no	59.7	57.8	58.7	58.2	59.7	58.4	58.9	58.7
logistic regression w. high-level features	yes	73.0	69.5	64.9	67.1	68.4	64.3	60.6	62.4
logistic regression w. bag of words	yes	76.7	74.1	70.2	72.1	74.2	71.1	66.9	69.0
RoBERTa	yes	85.4	83.4	83.4	83.4 (±0.1)	83.0	80.8	80.6	80.7 (±0.2)
BERTweet	yes	85.7	83.7	84.2	83.9 (±0.2)	83.4	81.2	81.3	81.3 (±0.3)

Table 4: Performance of different classifiers on 5-fold cross-validation on BRT-dataset.

Supervised Classifiers. We used **logistic regression** as a classifier for our high-level features (§4) and for bag of words. Furthermore, we considered the **two transformers** RoBERTa (Liu et al., 2019) and BERTweet (Nguyen et al., 2020). The latter is specifically designed for Twitter. Since we do not want our classifiers to overfit, we did not tune the hyperparameters on BRT but took the settings from Nguyen et al. (2020) which are generally considered effective for Twitter.[10]

Experimental Set-up. The supervised classifiers were evaluated via 5-fold cross-validation. We **created the folds in two different ways**: On the one hand, we randomly split the tweets of BRT into 5 different folds (**randomized folds**). These folds may produce training and test splits where tweets of the same timeline occur both in the training and the test set. Within a particular timeline, biographically relevant tweets may co-occur with particular topics or constructions. These co-occurrences are idiosyncratic to the user to whom that timeline belongs. However, they may not be representative of the classes to be predicted. Supervised classifiers may produce high classification scores by memorizing these user-specific artefacts.

As a realistic alternative, we consider a different set-up in which the tweets of a particular timeline are restricted to the same fold (**unseen timelines**). As a consequence, on this set-up classifiers are not rewarded for learning user-specific artefacts since the tweets in a test set will always originate from timelines not observed in the training data.

As far as the transformers are concerned, we further report the average result over 5 different runs (including standard deviation). All other classifiers produce deterministic output.

Results. The results of our evaluation are displayed in Table 4. Next to accuracy, we report macro-average precision, recall and F-score. The simple majority-class baseline is outperformed by every other baseline by a large degree in terms of F-score. This even includes the aspectual classifier which suggests that (despite the noise in the automatic analysis we already reported in §5) aspectual information is not totally uncorrelated to biographical relevance. The most effective baseline in terms

[10]The settings are: batch size=32; no. of epochs=30; learning rate=1e-05. (We use *roberta-large* and *bertweet-large*.)

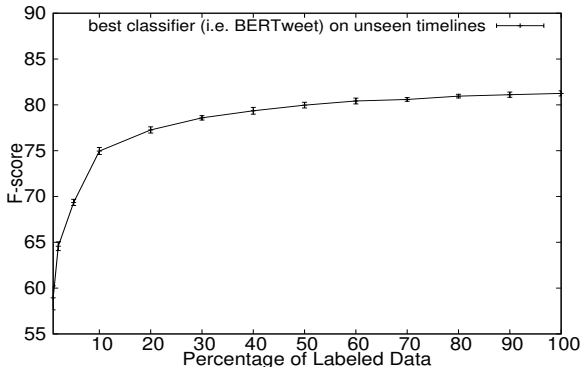

Figure 1: Learning curve on BRT-dataset.

of F-score is the classifier using meta-information. The supervised classifiers outperform the baselines. The most effective classifier is BERTweet, i.e. the language model that has been designed for Twitter.

Table 4 also contrasts the performance between *randomized folds* and *unseen timelines*. For all supervised classifiers, the evaluation based on randomized folds produces higher scores.[11] This confirms our assumption that randomizing folds produces overly optimistic results and that the classification on unseen timelines is a more realistic set-up and should be used in future evaluations.

Figure 1 shows the learning curve using our best performing classifier from Table 4 (i.e. BERTweet). With the full amount of the training data, the classifier seems to have reached a plateau. Therefore, we expect that additional training data will only result in marginal classification improvements.

7 Subtypes of Biographical Relevance

We extended our annotated BRT-dataset (§3) by an additional layer. For each biographically relevant tweet, we manually annotated the underlying subtype. Thus, we want to give an indication of what type of biographically relevant information can be found on Twitter. For the inventory of subtypes, we chose categories that are sufficiently distinctive and frequent on our data. Table 5 contains the distribution of the different subtypes as determined by our second layer of annotation. We also computed for each type the standard deviation across the different timelines. Only a high proportion with a low standard deviation indicates that the particular subtype is sufficiently frequent and can be consistently observed throughout the different timelines.

Table 5 shows that frequent biographically relevant subtypes are those that reveal likes of the author, other demographic information, family relations and instances that relate to job and education. From these subtypes, other demographic information has the lowest standard deviation, which means that this subtype can be more regularly observed across the different timelines at the same high frequency. Ideological views and habits occur similarly frequent but the standard deviation of ideological views is almost three times as high as the one of habits. This means that the latter occurs more regularly across the timelines.

8 Error Analysis

Despite the strong performance of our best classifier, i.e. BERTweet, we still observed a set of cases that were systematically classified incorrectly. A frequent error source stems from diverse forms of overgeneralization. While according to our annotation guidelines, desires and (dis)likes are considered biographically relevant, trivial desires (24) and (dis)likes (25) are not. Our classifier, however, fails to make this distinction. Similarly, while long-term health-conditions are biographically relevant (e.g. asthma, diabetes), common temporary illnesses are not (26). Our classifier also fails to make this distinction. Moreover, it is not able to appropriately take aspectual information into consideration. Episodic needs and desires (27) are thus erroneously considered biographically relevant.

(24) I just want some happiness *(irrelevant since everyone seeks happiness)*
(25) I Fuckin hate hiccups *(irrelevant since having a hiccup is generally considered an unpleasant experience)*
(26) Damn I have a cold again! *(irrelevant since a cold is a common illness everyone attracts now and then)*
(27) I need to get some sleep. *(irrelevant since this need/desire is only episodic)*

Finally, an obvious challenge are instances of implicit realizations of biographical relevance where there are no obvious lexical clues that suggest the presence of this category (28)-(29).

(28) Look at what arrived, my official Biden-Harris pins! *(inference: author is a Biden-Harris supporter)*
(29) He's fecking fat. Someone is feeding him. One of the neighbours. *(inference: author owns a cat)*

[11] It may come as a surprise that for the non-supervised classifiers the results between *randomized folds* and *unseen timelines* differ slightly. As a matter of fact, these classifiers do not change their prediction per tweet. These differences are the result of averaging over test folds whose size varies to a small degree between *unseen timelines* and *randomized folds*.

Subtype	Prototypical Example	Percent (std)
general likes	@USER I love Rage Against the Machine	20.6 (±12.4)
other demographic info.	I'm 6 foot 3. I stood next to a co worker today and I'm barely up to his neck	19.5 (±6.5)
job and education	@USER I'm a health worker There are thousands of us	18.4 (±13.3)
family relations	@USER Ooh my dead husband would have loved this	16.9 (±10.2)
general dislikes	i don't really like halloween...	16.5 (±14.1)
ideological views	@USER I think the future of humanity is very much in peril already	10.8 (±16.3)
habits	I refuse to see anyone before 11 Even though I am up before 9 and usually 8	10.4 (±5.3)
desires and needs	Want to move to a beach town in a few years	8.2 (±6.5)
food and drink preferences	@USER Ever eaten frogs They are amazing but we eat them too A tad more sentient than oysters	8.1 (±11.6)
health	@USER Omg I have a vasectomy! I'm not a male anymore apparently	5.4 (±4.2)
personality traits	@USER Yes I'm a sociable person. I miss conversation and dinner with someone who just might find me interesting...	3.9 (±2.8)
confidential information	Typing my surname earlier and it somehow suggests I'm 'Sillier' not Silbiger	2.8 (±3.8)
life events	@USER I moved to Georgia	1.6 (±1.5)
other	*not applicable*	6.2 (±4.4)

Table 5: Distribution of subtypes of biographical relevance. *(The same tweet may contain more than one subtype. Therefore, the sum of the percentages adds to more than 100%.)*

9 Conclusion

We presented a new dataset comprising tweets for the novel task of detecting biographically relevant utterances. Unlike previous research we do not restrict biographical relevance to a set of pre-defined relations. We showed that state-of-the-art classifiers are able to automatically detect such utterances even on tweets of unseen timelines. Our feature analysis using high-level features revealed the relatedness between linguistic features and biographical relevance (e.g. specific supersenses and POS-tags) while there is also a set of different meta features predictive for biographical irrelevance.

10 Ethical Considerations

The data we are going to make publicly available as part of this research will include tweets from specific timelines of Twitter. In order to protect the privacy rights of the authors, the usernames of the respective timelines have been anonymized by replacing each name by some generic ID. Moreover, we just release the IDs of the tweets contained in the timelines rather than the tweets themselves. The public release of such content as in the range of our dataset is also in accordance with the regulations of Twitter.

The manual annotation for our novel dataset was produced by a member of the department of one of the co-authors. The annotation was carried out as part of their regular work. Therefore, the work has been duly compensated.

The work presented in this paper addresses the task of extracting biographically relevant material published by ordinary people on Twitter. We are fully aware that research on this topic could be misused for the development of applications that we consider questionable from an ethical point of view. For instance, an application could be built that enables potential employers to do background checks on job seekers. Or, insurance companies could screen potential insurants before offering them a particular insurance. In §1, we acknowledged that potential of such research. However, we also made it clear that, in no way, this reflects the motivation of our research. On the contrary, we could envisage scenarios that would help users to protect the above misuse of their personal data. For instance, the functionality to identify biographically relevant information in user-generated content may be incorporated in an application that automatically anonymizes such sensitive text passages or, at least, warns users prior to posting such content that they are about to disclose sensitive material. Our work is also intended to raise public awareness of what contents can be found on Twitter and can be automatically and legally extracted.

11 Acknowledgements

The authors would like to thank Sybille Sornig for contributing to the manual annotation of our novel dataset.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of Association for Computational Linguistics: EMNLP 2020*, 1644–1650, Online.

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An Unsupervised Approach to Biography Production

Using Wikipedia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 807–815, Columbus, OH, USA.

Gerardo Canfora, Andrea Di Sorbo, Enrico Emanuele, Sara Forootani, and Corrado A. Visaggio. 2018. A Nlp-based Solution to Prevent from Privacy Leaks in Social Network Posts. In *Proceedings of the International Conference on Availability, Reliability and Security*, pages 1–6, Hamburg, Germany.

Paolo Cappellari, Soon Ae Chun, and Mark Perelman. 2017. A Tool for Automatic Assessment and Awareness of Privacy Disclosure. In *Proceedings of the Annual International Conference on Digital Government Research*, pages 586–587, Staten Island, NY, USA.

Thomas Dickinson, Miriam Fernández, Lisa A. Thomas, Paul Mulholland, Pam Briggs, and Harith Alani. 2015. Identifying Prominent Life Events on Twitter. In *Proceedings of the International Conference on Knowledge Capture, K-CAP*, pages 4:1–4:8, Palisades, NY, USA.

Atefeh Farzindar and Diana Inkpen. 2015. Natural Language Processing for Social Media. In *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1757–1768, Berlin, Germany.

Annemarie Friedrich and Manfred Pinkal. 2015. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2471–2481, Lisbon, Portugal.

Marcos Garcia and Pablo Gamallo. 2015. Exploring the effectiveness of linguistic knowledge for biographical relation extraction. *Natural Language Engineering*, 21(4):519–551.

Nikesh Garera and David Yarowsky. 2009. Structural, Transitive and Latent Models for Biographic Fact Extraction. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 300–308, Athens, Greece.

Ian Goodyer. 2001. *The depressed child and adolescent*, chapter Life events: Their nature and effects. Cambridge University Press.

Ozan Irsoy and Claire Cardie. 2017. Deep Recursive Neural Networks for Compositionality in Language. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pages 2096–2104, Montréal, Canada.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC KBP 2010 Knowledge Base Population Track. In *Proceedings of the Text Analytics Conference (TAC)*, Gaithersburg, MD, USA.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Jiwei Li and Claire Cardie. 2014. Timeline Generation: Tracking individuals on Twitter. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 643–652, Seoul, Korea.

Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014. Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1997–2007, Doha, Qatar.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4188–4203, Online.

Bing Liu. 2012. Sentiment Analysis and Opinion Mining. In *Synthesis Lectures on Human Lanugage Technologies*, volume 5, pages 1–167. Morgan & Claypool Publishers.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Delfina Malandrino, Andrea Petta, Vittorio Scarano, Luigi Serra, Raffaele Spinelli, and Balachander Krishnamurthy. 2013. Privacy Awareness about Information Leakage: Who knows what about me. In *Proceedings of the ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 279–284, Berlin, Germany.

Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In *Proceedings of the ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 1–12, Chicago, IL, USA.

Ben Medlock. 2006. An Introduction to NLP-based Textual Anonymisation. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 1051–1056, Genoa, Italy.

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244.

Saif Mohammad and Peter Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 39(3):555–590.

Jan Neerbek, Ira Assent, and Peter Dolog. 2017. TABOO: Detecting unstructured sensitive information using recursive neural networks. In *Proceedings of the IEEE International Conference on Data Engineering*, pages 1399–1400, Chania, Greece.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 9–14, Online.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. 2012. Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances. Technical Report CMU-ML-12-107, Machine Learning Department, Carnegie Mellon University.

Alistair Plum, Marcos Zampieri, Constantin Orasan, Eveline Wandl-Vogt, and Ruslan Mitkov. 2019. Large-scale Data Harvesting for Biographical Data. In *Proceedings of the Conference of Biographical Data in a Digital World*, pages 66–72, Varna, Bulgaria.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the Author Profiling Task at PAN 2013. In *Working Notes for CLEF 2013 Conference*, pages 352–365, Valencia, Spain.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying Latent User Attributes in Twitter. In *Proceedings of the International Workshop on Search and Mining User-Generated Contents (SMUC)*, pages 37–44, Toronto, ON, Canada.

Michaela Regneri, Alexander Koller, Josef Ruppenhofer, and Manfred Pinkal. 2011. Learning Script Participants from Unlabeled Data. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 463–470, Hissar, Bulgaria.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. *Recommender Systems Handbook*, chapter Recommender Systems: Introduction and Challenges. Springer.

Koustuv Saha, Jordyn Seybolt, Stephen M. Mattingly, Talayeh Aledavood, Chaitanya Konjeti, Gonzalo J. Martinez, Ted Grover, Gloria J. Mark, and Munmun De Choudhury. 2021. What Life Events are Disclosed on Social Media, How, When, and By Whom? In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22, Yokohama, Japan.

Satoshi Sekine and Javier Artiles. 2009. WePS2 Attribute Extraction Task. In *Proceedings of the Web People Search Evaluation Workshop (WePS)*.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 602–608, Minneapolis, MN, USA.

Xing Xie. 2010. Potential friend recommendation in online social network. In *Proceedings of the IEEE/ACM International Conference on Green Computing and Communications & International Conference on Cyber, Physical and Social Computing*, pages 831–835, Hangzhou, China.

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 387–390, Dublin, Ireland.

Liang Zhou, Miruna Ticrea, and Eduard Hovy. 2004. Multi-Document Biography Summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 434–441, Barcelona, Spain.