Measuring the Language of Self-Disclosure across Corpora

Anonymous ACL submission

Abstract

Being able to reliably estimate self-disclosure – a key component of friendship and intimacy – from language is important for many psychology studies. We build single-task models on five self-disclosure corpora, but find that these models generalize poorly; the within-domain accuracy of predicted message-level self-disclosure of the best-performing model (mean Pearson's r=0.69) is much higher than the respective across data set accuracy (mean Pearson's r=0.32), due to both variations in the corpora (e.g., medical vs. general topics) and labelling instructions (target variables: self-disclosure, emotional disclosure, intimacy). However, some lexical features, such as expression of negative emotions and use of first person personal pronouns such as 'I' reliably predict self-disclosure across corpora. We develop a multi-task model that improves results, with an average Pearson's r of 0.37 for out-of-corpora prediction.

1 Introduction

Interpersonal exchanges are a core component in human relationships. They are determined by intimacy, which in turn is characterized by the willingness of the involved parties to self-disclose (Rubin and Shenker, 1978). In general, self-disclosure can be defined as “revealing intimate information about one's self” (Derlega et al., 1993). Note that self-disclosure, which often involves revealing embarrassing facts about oneself that are considered violations of social norms ("I flunked my exam." or "I have a growth on my butt"), is different from revealing personally identifiable information (PII). Self-disclosure encompasses the sharing of thoughts, aspirations, feelings, likes and dislikes, while PII, such as date of birth or social security number, is used to unambiguously identify a person. Unlike self-disclosing, sharing PII does not necessarily suggest an intimate relationship between two people.

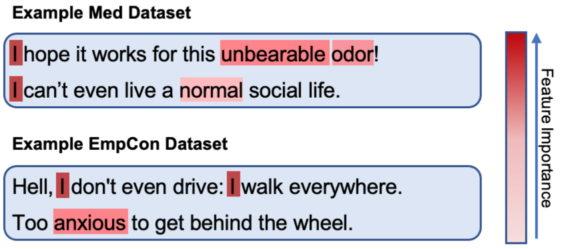

Example Med Dataset
I hope it works for this unbearable odor!
I can't even live a normal social life.
Example EmpCon Dataset
Hell, I don't even drive: I walk everywhere.
Too anxious to get behind the wheel.
Feature Importance

Figure 1: Two example sentences from the Med and the EmpCon data set. We predict the associated self-disclosure and highlight the most important features in both sentences.

NLP researchers have labeled a variety of data sets with self-disclosure or some approximation of self-disclosure such as "intimacy", which is more accurately viewed as being a property of the relationship between two people than of an utterance. In this paper, we build models to predict self-disclosure from text, and assess how well these models generalize across five different corpora. We find that they mostly generalize poorly, but that there are some reliable linguistic markers of self-disclosure.

We draw on multiple corpora labeled for self-disclosure: conversations from an online breast cancer support community (Wang et al., 2015); annotated conversational turns (Sedoc, N.D.); medical posts from patient.info and Reddit (Valizadeh et al., 2021) and posts from the r/OffMyChest and the r/CasualConversations subreddits (Jaidka et al., 2020). The labels on these data sets vary both in terms of how self-disclosure is defined, and in their scaling (e.g., 0/1 or 1-5 Likert scales), complicating the analysis.

Research questions We seek to determine

1. Which linguistic features predict self-disclosure in messages?
2. How well do language models trained on one data set predict self-disclosure in different corpora?

3. How to best build models that generalize self-disclosure across different corpora?

Better understanding the linguistic characteristics of self-disclosure is potentially useful in advising people on how to increase their self-disclosure, to increase intimacy and well-being. Self-disclosure is a key component of both romantic and platonic intimacy (Laurenceau et al., 1998) and is an indicator and influencing factor of self-esteem and well-being (Leung, 2002; Daley, 2010). Having more accurate models to identify self-disclosure in language will likely also support further research into the role of self-disclosure in areas ranging from depression treatment to friendship formation.

Contributions

1. We identify the linguistic correlates of self-disclosure, for example the expression of negative emotions and the use of first-person personal pronouns like 'I'.
2. We find that self-disclosure models generalize poorly across corpora due to the differences in their domains and labels.
3. We build a multi-task RoBERTa-based model, which gives the current state-of-the-art for the measure of self-disclosure across multiple corpora.

We make the code and data for all experiments available on GitHub.[1]

2 Background and Related Work

People reveal information about themselves to form and maintain personal relationships (Joinson and Paine, 2007). As an essential part of interpersonal communication, self-disclosure can have both positive and negative effects on the person disclosing, which are reinforced in an online environment. Risks resulting from revealing private information can encompass a loss of privacy (Haimson et al., 2015; Vitak and Kim, 2014), a negative impact on identity and self-presentation (Morris and Millen, 2007), as well as negative consequences caused by context collapse, i.e. the disclosure to an unintended audience, that is especially prevalent in social media environments (Farnham and Churchill, 2011). On the other hand, disclosing private information can lead to increased social expression, social validation and perceived intrinsic rewards (Pennebaker, 1993; Goldfried et al., 2003).

Self-disclosure can be influenced by a variety of factors including anonymity, cultural norms, personality, loyalty and mutual trust (Postmes et al., 2001; Laursen, 1993). These have an impact on the risk/benefit dynamic in revealing personal information online. Bazarova and Choi (2014) have formulated a functional model of self-disclosure to capture these conflicting dynamics and allow for a more holistic understanding of self-disclosure by showing how people try to maximize their benefits when disclosing private information.

Self-disclosure is a determining factor in the level of intimacy between people, which in turn defines the quality of relationships. On an individual level, it has been shown that intimate relationships are an important resource for inter- and intrapersonal growth (Buhrmester, 1990). They strengthen a person's sense of belonging and self-worth (Rawlins, 2017) and provide a source of emotional support as well as a safe space for self-exploration (Buhrmester, 1990; Parker and Gottman, 1989). Through these mechanisms, self-disclosure can positively influence a person's mental health (Stiles, 1987), improving their feeling of connectedness to others, a primary human need (Ryan and Deci, 2000). For example, Buhrmester (1990) showed that intimate relationships, which are dependent on self-disclosure, lead to better competence, sociability and self-esteem as well as less self-reported depression and anxiousness, compared to reference groups with less intimate connections.

The steady rise of social media usage led to an increase in the availability of publicly disclosed 'private' information. This is especially interesting given that self-disclosure has been found to be higher online compared to face-to-face communication (Tidwell and Walther, 2002; Joinson and Paine, 2007), partially because sharing to larger audiences is facilitated in an online context (Bazarova, 2012). In the light of these developments, we use social networking sites (SNS) data to identify and subsequently predict self-disclosure in online posts. In previous works, self-disclosure was predicted in different contexts using unsupervised, semi-supervised and supervised models. Blose et al. (2020) used unsupervised learning to detect the voluntary disclosure of private information in Tweets. They investigated how self-disclosure was impacted due to the COVID-19 pandemic and found a significant shift towards support-seeking and supportiveness. In addition, Bak et al. (2014)

[1] https://github.com/ Code and data will be released upon acceptance.

developed a semi-supervised self-disclosure topic model to automatically detect self-disclosure in tweets, with the aim of analyzing its effects on subsequent conversations. They find a significant positive correlation between self-disclosure and conversation length as well as frequency. Furthermore, Yang et al. (2017) investigated how publicness influences self-disclosure in health support groups by applying a supervised model based on the Linguistic Inquiry and Word Count (LIWC) as well as other linguistic features and word embeddings to assess the level of positive and negative self-disclosure. Considering the broader concept of intimacy, Pei and Jurgens (2020) designed a computational framework to study the expression of intimacy in questions. They predicted intimacy using a semi-supervised model, showing that it is an impactful dimension in language that is influenced by social settings.

Our study differs in that we aim to understand self-disclosure across different platforms and contexts. We focus on the specific prediction of self-disclosure to contribute to previous efforts, for example by (Preoţiuc-Pietro et al., 2015), to assess well-being and mental health from social data. As such, we are not limited to one SNS but rather aim to develop a supervised model that generalises across multiple platforms. We further compare the performance of RoBERTa-, LIWC-, LDA- and EmoLex-based models to show which linguistic features are predictive of self-disclosure. Finally, given that we find that single-task models are insufficient, we develop a multi-task model across all available data sets to assess self-disclosure. This is an innovative approach that has not yet been pursued in this realm to the best of our knowledge.

3 Data Sets

To develop a general model to detect the degree of self-disclosure in messages, we gathered five data sets, trained models on them, and tested the performance of these models across all data sets.

The available data sets offer a challenge in that they all have different labels, including ’self-disclosure’, ’intimacy’, and ’emotional disclosure’. These labels differ both in the instructions provided to the annotators (there is no consistent definition of self-disclosure used in computational linguistics) and in their scales. The fact that some labels are binary and others are on 1-to-3, 0-to-5, or 1-to-7 Likert scales complicates the analysis. We thus evaluate the accuracy of models by looking at the correlation of the prediction with the true label, allowing us to see e.g. how accurately a prediction of a 1-to-5 label estimates a binary label.

Data Set	Source	Size
OnSup	online support forum	1 000
OffChe	Reddit	12 860
Int	Reddit	2 387
EmpCon	conversations by MTurk workers	5 820
Med	patient.info	6 417

Table 1: Overview of the data sets considered.

Online Support data set (OnSup) The OnSup data set was collected by Wang et al. (2015) from discussion boards and chat rooms of an online breast cancer support community. The authors randomly selected 1000 exchanges, of which the thread-starting messages were each manually labeled by ten Amazon Mechanical Turk (MTurk) workers for positive and negative self-disclosure, among other categories. Self-disclosure in this context was defined as the "the extent to which the writer has discussed her feelings and emotions with others, such as happiness, fears, sadness, and anger." Given examples for positive self-disclosure included phrases like "Now that chemo is done, I find myself waking up in the morning feeling a huge burden has been lifted from my shoulders." and "I am freaked out after reading my mammogram report." for negative self-disclosure. The individual ratings, ranging from 1 (no self-disclosure) to 7 (a great deal of self-disclosure) were combined by taking the workers’ average. To allow for the comparison across our data sets, we introduced a general self-disclosure indicator for this data set by adding together the negative and positive self-disclosure scores from the initial study.

Empathic Conversations data set (EmpCon) The EmpCon data set by Sedoc (N.D.) contains 5819 conversational turns, where each turn has been labeled by four MTurk workers for empathy, emotion, emotional polarity and self-disclosure. The instructions the annotators were given with regards to self-disclosure included the following Human Intelligence Task (HIT): "When judging self-disclosure, think: Did this make you know the writer of the statement better?". The workers labeled the degree to which they agreed with this notion on a scale from 1 to 3, where 1 corresponded to 'Not at all' to 3 'A lot'.

Medical data set (Med) The Med data set by Valizadeh et al. (2021) contains online conversations from randomly-selected forums on patient.info and other online platforms, filtered for medical keywords and hashtags. Each message was labeled for medical self-disclosure. The assigned labels ranged from 0 to 5, where 0 corresponded to 'no self-disclosure' and 5 indicated 'high self-disclosure'. The label '5' was given for instances were the post writer specifically mentioned that he/she was diagnosed with a specific illness, was taking specific medication, had undergone surgery or was about to have one, or other cases of disclosing specific medical indicators.

OffMyChest data set (OffChe) Jaidka et al. (2020) collected the OffMyChest conversations data set by letting 12860 Reddit top comments of the top posts from the r/OffMyChest and the r/CasualConversations subreddits be labeled for emotional disclosure on a binary scale. The latter was defined as comments mentioning the authors personal feelings e.g. "My only concern was for my son." and "My heart is breaking for you.".

Intimacy data set (Int) Compared to the previous four data sets, the fifth one we're taking into consideration is focused on 2397 questions drawn from question-centered subreddits such as r/AskReddit. However, instead of being labeled for self-disclosure, the questions were evaluated for intimacy, which was defined by the authors Pei and Jurgens (2020) as "how an individual relates to their audience in their perceived interdependence, warmth, and willingness to personally share". They employed a best-worst-scaling for labeling by showing annotators a tuple of four questions, among which the least and most intimate question should be identified. That way, five pairwise comparisons were obtained per tuple that were used as part of a Luce Spectral Ranking (Maystre and Grossglauser, 2015) to infer a continuous latent intimacy score on a scale from -1 (least intimate) to 1 (most intimate).

4 Features

Each of the above-mentioned data sets have been used to train discriminative, supervised machine learning models to correlate linguistic characteristics with the perceived presence of self-disclosure. In this section, we present the features we took into consideration.

N-gram distributions We tokenized the texts using the Happier Fun Tokenizer (Schwartz et al., 2017) and extracted uni-, bi- and trigrams.

LIWC The theory-based LIWC lexicon (Pennebaker et al., 2007) is widely used to analyze the usage of word semantic categories within text. It contains 73 categories ranging from parts of speech to emotions and cognitive styles, including personal pronouns such as 'I', which have been shown to be related to self-disclosure, and collections of words for positive and negative emotions (called POSEMO and NEGEMO respectively). LIWC word frequencies capture emotions well, and thus are expected to correlate with self-disclosure, since emotions are more associated with self-disclosure than facts.

LDA topics Given that data-driven topics tend to be more representative of online posts, we also used Latent Dirichlet Allocation (LDA) Facebook topics. This is a normalized frequency distribution of 2000 topics based on a Facebook corpus with approximately 18 million posts obtained from the Differential Language Analysis ToolKit (DLATK) repository (Schwartz et al., 2013). We used these topics to uncover hidden topics as well as words that represent these topics in the data sets.

Emotion lexica High self-disclosure statements tend to be more emotional. In addition to the emotion-related categories in LIWC, we used the NRC EmoLex lexicon which has 14182 manually labeled entries for the emotions 'anger', 'anticipation', 'disgust', 'fear', 'happiness', 'sadness', 'surprise' and 'trust' as well as 'positive' and 'negative prevalence'.

RoBERTa embeddings Finally, we considered word embeddings, i.e. real-numbered vectors mapped from words or phrases representing their distributional semantic meaning, to obtain a conceptualized token embedding. In this context, RoBERTa, a bi-directional transformer (Liu et al., 2019), was used for classification using sentence representations obtained from the model. Specifically, we used RoBERTa embeddings as features in our proposed models.

5 Models

5.1 Single-Task Models

A five-fold cross-validated Ridge regression with the data set-specific target variables was trained separately on 1-to-3 grams, LIWC, LDA and EmoLex topics, as well as RoBERTa embeddings for each of the target data sets. The alpha values used can be found in Table 8 in the appendix. We subsequently used the best-performing model for each data set to predict self-disclosure on the other data sets to assess the across-data set accuracy of the single-task models.

5.2 Multi-Task Models

In addition to the described single-task models, we developed models based on LIWC and RoBERTa features in which multiple tasks, i.e. the prediction of the different notions of self-disclosure across the available data sets, were learned simultaneously. We expected that multi-task learning would improve the results obtained by the single-task model. However, compared to standard multi-task learning, we faced the issue that each of the data sets had different outcomes on different scales. Thus, in contrast to standard multi-task learning, where outcomes for all tasks are available for each instance, we were missing 4/5th of the labels for each observation.

Estimating a model across the multiple data sets thus requires handling the fact that the labels on each data set are different – and are on different scales. One option to handle this would be to translate all the labels to lie on the same range. This, however, assumes that a linear transformation would suffice, and that the correct transformation could be found. Instead, we build a single neural net that takes in an embedding of the post, and outputs predictions for all of the labels. Given the relatively small training sets, we used a neural network with one single-dimensional hidden layer. The output of that hidden layer can be viewed as a latent variable capturing self-disclosure, which is then transformed to yield each of the actual self-disclosure labels. For any given observation, only one label is observed, so that training loss is estimated as the sum over the training data (e.g., all observations in three of the four data sets) of the loss on the label that is present for that observation. Note that the loss is the squared error for continuous labels and the cross entropy for discrete labels. The labels for each continuous data set were normalized to zero mean and unit variance to put all losses on a similar scale.

Since we are interested in the statistical similarity between the labels of the different data sets, Pearson's r values between the single-dimensional latent variable and the holdout data set labels are reported. Networks with and without a sigmoid activation after the hidden dimension were explored with the latter found to be more effective. Hyperparameters and optimization details can be found in Tables 9, 10, and 11 in the appendix.

6 Results

We now discuss the quantitative results and their implications. Since we found that the Int data set does not generalize well (due to the fact that it mainly consists of questions), we focus on the four remaining data sets and only report the Int results in the appendix.

6.1 General Model to Predict Self-Disclosure

We computed both the single-task and multi-task models for the different data sets. Starting with the former, we calculated the within-data set Pearson's r based on a Ridge regression for different feature sets for all considered data sets:

Model	Emp-Con	OnSup	Med	Off-Che
Ngrams	0.64	0.53	0.61	0.17
LIWC	0.64	0.66	0.64	0.29
LDA	0.57	0.22	0.62	0.41
Emo	0.32	0.25	0.19	0.10
RoBERTa	**0.73**	**0.72**	**0.85**	**0.47**

Table 2: Prediction performance for self-disclosure models (captured by Pearson's r) within data sets, averaged over a five-fold cross-validation.

Table 2 shows that the in-domain prediction of self-disclosure was generally most accurate with RoBERTa embeddings. We therefore used these RoBERTa embedding-based models to calculate the cross-data set performance, shown in Table 3.

The across-data set Pearson's r ranges from 0.16 to 0.48, with an average of 0.32, a significant drop compared to the best-performing (i.e. RoBERTa) within-data-set average r of 0.69.[2] Looking at the

[2] While the Pearson's r scores for the considered models are low compared to many results in computational linguistics, which are between 0.8 and 0.95 for problems like POS tagging, they are in line with most predictions of psychological

	Emp-Con	**OnSup**	**Med**	**Off-Che**
EmpCon	(0.73)	**0.42**	**0.48**	**0.21**
OnSup	**0.44**	(0.72)	0.35	0.16
Med	0.19	0.28	(0.85)	0.17
OffChe	0.34	0.41	0.44	(0.47)
Avg	0.32	0.37	0.42	0.18

Table 3: Across-data-set prediction results (Pearson's r) for self-disclosure, using RoBERTa embeddings. The first column shows the data set the model has been trained on, the first row the data set it has been tested on. The diagonals are within data set cross-validation accuracies. The last row shows the average of the Pearson's r values for the respective column, excluding the within-data-set accuracy reported in brackets.

individual across-data set Pearson's r values, we find that the EmpCon data set, consisting of labeled conversation turns, performs reasonably well on the OnSup samples, most likely because both data sets resemble more structured conversations instead of single independent posts.

Predictive accuracies for the linear multi-task model are presented in Table 4. As expected, single-task models performed best on the same corpus that they were trained on. On average, the out-of-task multi-task models outperformed the across-data set single task models (single-task across data set average: r=0.32, linear multi-task average: r=0.37). We found that a multi-task model trained on the EmpCon, the OnSup, and the OffChe data sets performed best. This is in line with our expectations, since these three data sets are less domain-specific than the Med data set and hence, generalize better.

Both the linear and the nonlinear multi-task models based on LIWC features performed worse than the multi-task models based on RoBERTa embeddings, which is why we only report them in the appendix. Given these results, we recommend a linear multi-task model based on all data sets we considered to predict self-disclosure on a message level. The corresponding code and model will be made available upon publication.

Target Data Set	**Linear**
EmpCon	0.37
OnSup	0.42
Med	0.46
OffChe	0.24
Avg	0.37

Table 4: Prediction results (Pearson's r) for linear multi-task models based on RoBERTa embeddings. The first column is the target data set for the respective model that was trained on the remaining three data sets. The nonlinear results are similar and reported in the appendix.

6.2 Linguistic Features Predictive of Self-Disclosure

We found a strong positive correlation between use of the personal pronoun 'I' (as captured by the LIWC category 'I') and self-disclosure across all data sets, and a similarly strong negative correlation between the use of 'you' and self-disclosure. This is to be expected; there should be more self-disclosure when talking about oneself than when talking about the person you are talking to. Furthermore, interrogatives, i.e. question words, are negatively correlated with self-disclosure across all considered data sets. Asking questions is low self-disclosure. It is worth noting that the signal for predicting self-disclosure is spread over many more categories of words; simply using 'I', 'you' and questions is insufficient to build an accurate model.

Topic	**Emp-Con**	**OnSup**	**Med**	**Off-Che**
I	**0.35**	**0.36**	**0.44**	**0.16**
THEY	0.07	-	-	-0.03
SHEHE	0.07	0.12	-0.06	-
WE	0.06	-	-0.09	-
YOU	-0.29	-0.13	-0.40	-0.05

Table 5: LIWC-based classifier accuracy: Pearson's r of the linguistic topics for all data sets at the $p < 0.01$ level. A hyphen indicates that the respective category was not significant.

Positive emotions correlate much more weakly with self-disclosure than negative emotions do, as shown by both LIWC emotion (Table 6) and EmoLex topics (Table 12 in the appendix). This is consistent with the norm violation notion of self-

constructs, where r values of about 0.3 to 0.4 are the norm due to the wider range of unobserved factors influencing them. In this context, average Pearson's r scores, especially across data sets, of 0.3 on average indicate a significant predictive signal.

Topic	Emp-Con	OnSup	Med	Off-Che
NEGEMO	**0.24**	**0.45**	0.07	**0.12**
SAD	0.13	0.18	-	0.08
ANX	0.11	0.38	**0.08**	0.04
ANGER	0.11	0.22	-	0.09
POSEMO	-0.05	-0.14	-0.21	0.12

Table 6: LIWC-based classifier accuracy: Pearson's r of the emotion topics for all data sets at the $p < 0.001$ level. A hyphen indicates that the respective category wasn't significant.

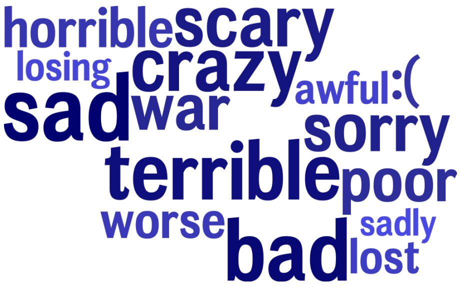


Figure 2: Sample correlation of LIWC NEGEMO words with self-disclosure based on the *EmpCon data set*, depicted as LIWC topic cloud. The size of each category is proportional to its correlation with the considered target label. Correlations are significant at $p < 0.01$.

disclosure mentioned in the introduction; Due to socio-cultural norms, interpersonal interactions are constrained with regards to acceptable or desired behavior (Allan, 1993). Disclosure of personal, negative emotions poses a higher risk in that it is a violation of norms (Caltabiano and Smithson, 1983), while the disclosure of positive information such as accomplishments is more normative. Thus positive emotions (POSEMO) correlate predominantly negatively with self-disclosure across the data sets.

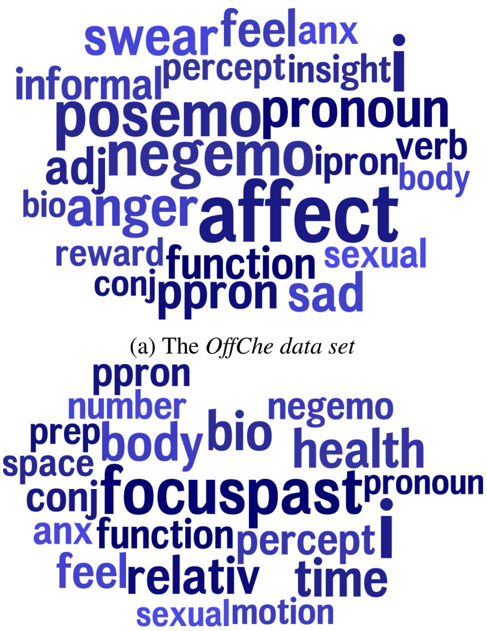
6.3 Generalization across Different Corpora

A key issue in building self-disclosure models is the differing labels based on differing definitions of self-disclosure across the data sets considered. (This is a common problem in computational social science, where constructs such as "happy" or "liberal" are often measured using widely different measures, see Casper et al. (2018) for more information). We handled this by introducing the multi-task model described above. In this section, we discuss differences in the predictive linguistic markers found across the considered data sets, and in the ability of our models to predict self-disclosure.

We found that self-disclosure models based on the Int data set generalized extremely poorly (average Pearson's r=0.14, see Table 16 in the appendix). The Int collection is not representative of self-disclosure because it only includes questions, which only obliquely reveal information about the person asking them. As mentioned above, we thus only reported the results from the Int data set in the appendix, and focused on the remaining data sets in our analysis.

The Med data set is also qualitatively different from the other data sets in that it is domain-specific. Revealing medical information is often particularly self-disclosing. Many medical conditions can be embarrassing to disclose to strangers because information related to illness tend to be negative and potentially embarrassing, hence disclosing medical information is norm-violating. Interestingly, negative emotions in a medical context are not as predictive of self-disclosure as in more general data sets like the other three considered in this paper. A possible explanation for these deviations in posts related to the medical domain is that norms in this context differ from general norms: Strong emotions like anger, disgust or sadness are less prevalent when talking about medical diagnoses and indicators, while the medical information itself is already considered a highly personal information, leading to a higher self-disclosure scores without the presence of negative emotions. This is supported by the results in Table 7: Compared to the other data sets, we find that the BIO and HEALTH categories show a stronger positive correlation to self-disclosure in the Med data set than in the other corpora. Interestingly, strong emotions like anger or anxiety tend to be less prevalent in this domain-specific data set, too, presumably for the above-mentioned reasons.

We further observe that the OffChe data set has less overall explanatory power within-data-set than the other data sets, but shows a relatively stable across-data-set performance. This is possibly because the OffChe data set has more than 12,000 data points, allowing for a better generalization, but at the same time it has lower internal predictive accuracy because self-disclosure was only measured on a binary scale. The per-message signal is thus weaker for OffChe data points than for the other data sets for which the target variable was measured on a continuous scale. This is confirmed by the results in Table 7, where all LIWC categories



(a) The *OffChe data set*

(b) The *Med data set*

Figure 3: Correlation of LIWC categories with self-disclosure in (a) the *OffChe data set* and (b) the *Med data set*. The size of each category name is proportional to its correlation with the self-disclosure label. Correlations are significant at $p < 0.01$.

are significantly less predictive of the OffChe data than for the other data sets.

Topic	**Emp-Con**	**OnSup**	**Med**	**Off-Che**
FUNCTION	0.36	0.30	0.09	0.05
I	0.35	0.36	0.44	0.18
NEGEMO	0.24	0.45	0.07	0.12
CONJ	0.22	0.13	0.10	0.02
PPRON	0.17	0.37	0.08	0.07
BIO	0.14	-	0.20	0.02
HEALTH	0.12	-	0.19	-
ANX	0.11	0.38	0.08	0.04
ANGER	0.11	0.22	-	0.09
FOCPAST	0.03	0.12	0.31	-
POSEMO	-0.05	0.08	-0.21	0.12
AFFECT	0.12	0.13	-0.14	0.18

Table 7: Top 5 significant, positively correlated LIWC categories per data set and corresponding Pearson r's for all data sets, sorted by decreasing values in the EmpCon data set.

7 Limitations & Ethical Considerations

Several limitations of our study should be taken into account when considering results in a wider context. Firstly, we have not studied how self-disclosure prediction differs among different cultures, genders and races. Specifically, it is unclear how well our recommended general self-disclosure model applies to specific subgroups. For example, women tend to self-disclose more and express more emotional content than men. Whether this suggests that different models of self-disclosure would be helpful for men and women is less clear. Similarly, the amount of self-disclosure varies widely across settings and cultures. How this affects models is similarly unclear. These variations should be studied in a subsequent research project. Secondly, our training corpora included mostly native English speakers and hence might not generalize well to non-native speakers. Finally, self-disclosure detection could be used for unethical targeting, e.g. in the context of insurance companies who want to discriminate on prices for people who don't self-disclose much, given that self-disclosure can influence relationships and subsequently the mental health of a person. The application of our model for such usages is strongly advised against.

8 Conclusion

Self-disclosure is a determining factor of the quality of interpersonal relationships, where closer friendships include more self-disclosure (Rubin and Shenker, 1978). Furthermore, the amount of self-disclosure on a platform should also strongly affect how much information can be extracted about personality and emotion from language written on that platform; Linkedin, for example, should show less self-disclosure than Facebook. Motivated by these observations, we studied to what extent self-disclosure can be predicted by looking at lexical features. Many aspects of language indicate self-disclosure. The expression of negative emotions and the use of first person pronouns are particularly predictive. Models trained on different data sets with different annotations of self-disclosure generalize poorly across corpora. Our best performing model, a RoBERTa-based linear multi-task model trained on on all our data sets, will be made available upon publication of this paper.

References

Graham Allan. 1993. Social structure and relationships. *Social context and relationships*, 3:1–25.

JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014.

Self-disclosure topic model for classifying and analyzing twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1986–1996.

Natalya N Bazarova. 2012. Public intimacy: Disclosure interpretation and social judgments on facebook. *Journal of Communication*, 62(5):815–832.

Natalya N Bazarova and Yoon Hyung Choi. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication*, 64(4):635–657.

Taylor Blose, Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2020. Privacy in crisis: A study of self-disclosure during the coronavirus pandemic. *arXiv preprint arXiv:2004.09717*.

Duane Buhrmester. 1990. Intimacy of friendship, interpersonal competence, and adjustment during preadolescence and adolescence. *Child development*, 61(4):1101–1111.

Marie Louise Caltabiano and Michael Smithson. 1983. Variables affecting the perception of self-disclosure appropriateness. *The Journal of Social Psychology*, 120(1):119–128.

Wendy J Casper, Hoda Vaziri, Julie Holliday Wayne, Sara DeHauw, and Jeffrey Greenhaus. 2018. The jingle-jangle of work–nonwork balance: A comprehensive and meta-analytic review of its meaning and measurement. *Journal of Applied Psychology*, 103(2):182.

Andrea Daley. 2010. Being recognized, accepted, and affirmed: Self-disclosure of lesbian/queer sexuality within psychiatric and mental health service settings. *Social Work in Mental Health*, 8(4):336–355.

Valerian J Derlega, Sandra Metts, Sandra Petronio, and Stephen T Margulis. 1993. *Self-disclosure.* Sage Publications, Inc.

Shelly D Farnham and Elizabeth F Churchill. 2011. Faceted identity, faceted lives: social and technical issues with being yourself online. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 359–368.

Marvin R Goldfried, Lisa A Burckell, and Catherine Eubanks-Carter. 2003. Therapist self-disclosure in cognitive-behavior therapy. *Journal of clinical psychology*, 59(5):555–568.

Oliver L Haimson, Jed R Brubaker, Lynn Dombrowski, and Gillian R Hayes. 2015. Disclosure, stress, and support during gender transition on facebook. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 1176–1190.

Kokil Jaidka, Iknoor Singh, Jiahui Lu, Niyati Chhaya, and Lyle Ungar. 2020. A report of the CL-Aff OffMyChest Shared Task: Modeling Supportiveness and Disclosure. In *Proceedings of the AAAI-20 Workshop on Affective Content Analysis*, New York, USA. AAAI.

Adam N Joinson and Carina B Paine. 2007. Self-disclosure, privacy and the internet. *The Oxford handbook of Internet psychology*, 2374252.

Jean-Philippe Laurenceau, Lisa Feldman Barrett, and Paula R Pietromonaco. 1998. Intimacy as an interpersonal process: The importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges. *Journal of personality and social psychology*, 74(5):1238.

Brett Paul Laursen. 1993. *Close friendships in adolescence.* Jossey-Bass, Inc.

Louis Leung. 2002. Loneliness, self-disclosure, and icq (" i seek you") use. *CyberPsychology & Behavior*, 5(3):241–251.

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*.

Lucas Maystre and Matthias Grossglauser. 2015. Fast and accurate inference of plackett–luce models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Joan Morris and David R Millen. 2007. Identity management: multiple presentations of self in facebook. In *Proceedings of the 2007 international ACM conference on Supporting group work*.

Jeffrey G Parker and John M Gottman. 1989. *Social and emotional development in a relational context: Friendship interaction from early childhood to adolescence.* John Wiley & Sons.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.

James W Pennebaker. 1993. Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour research and therapy*, 31(6):539–548.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135.

Tom Postmes, Russell Spears, Khaled Sakhel, and Daphne De Groot. 2001. Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*, 27(10):1243–1254.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717.

William Rawlins. 2017. *Friendship matters*. Routledge.

Zick Rubin and Stephen Shenker. 1978. Friendship, proximity, and self-disclosure 1. *Journal of Personality*, 46(1):1–22.

Richard M Ryan and Edward L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations*, pages 55–60.

Joao Sedoc. N.D. Empathic conversations: A multi-level dataset of contextualized conversations. *Note: not published yet*.

William B Stiles. 1987. I have to talk to somebody. In *Self-disclosure*, pages 257–282. Springer.

Lisa Collins Tidwell and Joseph B Walther. 2002. Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: Getting to know one another a bit at a time. *Human communication research*, 28(3):317–348.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4398–4408, Online. Association for Computational Linguistics.

Jessica Vitak and Jinyoung Kim. 2014. "you can’t block people offline" examining how facebook’s affordances shape the disclosure process. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 461–474.

Yi-Chia Wang, Robert E Kraut, and John M Levine. 2015. Eliciting and receiving online support: using computer-aided content analysis to examine the dynamics of online social support. *Journal of medical Internet research*, 17(4):e99.

Diyi Yang, Zheng Yao, and Robert Kraut. 2017. Self-disclosure and channel difference in online health support groups. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

A Appendix

A.1 Model Architectures

This section includes additional information about our single- and multi-task model architectures.

A.1.1 Single-task Model

In Table 11, we report the alpha values used in the single-task within-data set models. They were determined by a grid search over [0.0001, 0.001, 0.01, 1, 10, 100, 1000].

Topic	**Emp-Con**	**On-Sup**	**Med**	**Off-Che**
Ngrams	0.01	0.01	0.01	1
LIWC	0.01	0.01	1	1
LDA	0.01	0.01	0.01	0.01
Emo	0.01	0.01	0.01	1
ROB	100	100	10	100

Table 8: Alpha values for within-data set, single-task self-disclosure models.

A.1.2 Multi-task Model

For the multi-task models, we computed the optimal number of epochs for each considered learning rate ([1e-3, 1e-4, 1e-5]), where the learning rate was decreased by a factor of 10 when validation loss was static for 25 epochs. Afterwards, we performed for each target data set a five-fold cross-validation on the combined task of the three remaining data sets. Our batch size was 512 and we applied Adam optimization. If a batch was missing one of the data sets, it was skipped, so each batch contained all tasks. Heterogeneous batches were normalized by the number of examples in a batch and labels were normalized to the 0-1 range if they were continuous. As loss functions, we used the Mean Squared Error for continuous labels and the Binary Cross Entropy loss for discrete labels. The training was stopped when the learning rate reached 1e-6. The weighting was done equally by task. Note that in our multi-task training, almost all outputs were missing, since we didn't have all the different self-disclosure labels across all data sets but rather one specific one per data set.

For the initial **linear multi-task model**, we used a weight decay of 1.0 and a maximum learning rate of 1e-1. We let the model with the architecture shown in Table 9 train for 500 epochs.

Architecture Linear Model
Linear Layer from feature space to single dimension
Linear Layer from single dimension to output dimension (= number of tasks)

Table 9: Linear multi-task model architecture.

In addition, we found that the **nonlinear multi-task models** described in Table 10 turned out to be optimal for the **RoBERTa features**. This model trained for 300 epochs with a maximum learning rate of 2e-1 and a weight decay of 0.001.

Architecture Nonlinear RoBERTa Model
Dropout Layer with p=0.2
Linear Layer from feature space to 10 dimensions
Dropout Layer with p = 0.2
Batch Normalization Layer
Sigmoidal Activation
Linear Layer from 10 dimensions to single dimension
Batch Normalization Layer
Sigmoidal Activation
Linear Layer from single dimension to output dimension = number of tasks

Table 10: Nonlinear RoBERTa multi-task model architecture.

Finally, the **nonlinear multi-task model** reported in Table 11 was optimal for the **LIWC features**. It was trained over 300 epochs with a maximum learning rate of 5e-1 and a weight decay of 0.05.

Architecture Nonlinear LIWC Model
Dropout Layer with p=0.2
Batch Normalization Layer
Linear Layer from feature space to single dimension
Sigmoidal Activation
Batch Normalization Layer
Linear Layer from single dimension to output dimension = number of tasks

Table 11: Nonlinear LIWC multi-task model architecture.

A.2 Additional Results

In this section, we show additional results from our analysis, including the EmoLex classifier, the single-task results for the Int data set (both within- and across data set), the linear and nonlinear multi-task models based on LIWC as well as the nonlinear multi-task model based on RoBERTa embeddings.

A.2.1 EmoLex-based Classifier

Table 12 shows the results for the EmoLex-based classifier. Since they were in line with the emotion-related LIWC categories, we only reported the latter in the main text.

Topic	Emp-Con	On-Sup	Med	Off-Che
Anger	0.18	0.21	0.05	0.07
Anticip	-0.23	-	-0.08	0.04
Disgust	0.16	0.12	0.09	0.07
Fear	0.15	0.20	0.10	0.03
Joy	0.03	-0.09	-0.13	0.07
Sadness	0.17	0.26	0.10	0.05
Surprise	-	-	-0.08	0.04
Trust	0.04	-	-0.09	0.04
Positive	0.07	-0.10	-0.16	0.05
Negative	0.21	0.31	0.11	0.07

Table 12: Summary of the EmoLex-based classifier showing Pearson's r of the emotion topics for all data sets at p < 0.001. A hyphen indicates that the respective category wasn't significant.

A.2.2 Int Data Set

As discussed in the main text, we omitted the predictions from the Int data set since the corpus wasn't representative for our purposes as it only contained questions. In Tables 13 and 14, the key linguistic characteristics of the Int data set are shown.

Topic	Pearson's r
I	-
THEY	-
SHEHE	0.07
WE	-0.07
YOU	0.46

Table 13: LIWC-based classifier accuracy: Pearson's r of the pronoun topics for the Int data set at the p < 0.01 level. A hyphen indicates that the respective category wasn't significant.

Topic	Pearson's r
SAD	0.06
ANX	0.14
ANGER	0.07
POSEMO	0.05
NEGEMO	0.18

Table 14: LIWC-based classifier, reported as Pearson's r of the emotion topics for the Int data set at the p < 0.01 level. A hyphen indicates that the respective category wasn't significant.

In Table 15, we present the within-data set results for models based on the Int data set, averaged over a five-fold cross validation.

Model	Pearson's r
Ngrams	0.66
LIWC	0.64
LDA	0.55
EmoLex	0.08
RoBERTa	0.80
Avg	0.55

Table 15: Prediction performance for self-disclosure models based on the Int data set (captured by Pearson's r) within-data set, averaged over a five-fold cross-validation.

Table 16, on the other hand, shows the across-data set results for the best-performing within-data set Int model, i.e. the RoBERTa model, applied to all other considered data sets.

Data Set	Pearson's r
EmpCon	0.07
OnSup	0.29
Med	0.04
OffChe	0.16
Avg	0.14

Table 16: Prediction performance for Int self-disclosure RoBERTa model (captured by Pearson's r) across-data set averaged over a five-fold cross-validation.

A.2.3 Multi-task Models

In this section, we report additional multi-task models. Table 17 shows the results for the **RoBERTa-based nonlinear multi-task model**.

Target Data Set	**Pearson's r**
EmpCon	0.45
OnSup	0.29
Med	0.34
OffChe	0.22
Avg	0.33

Table 17: Prediction results (Pearson's r) for nonlinear multi-task models based on RoBERTa embeddings. The first column is the target data set for the respective model that was trained on the remaining three data sets.

Table 18 shows the results for the **LIWC-based nonlinear multi-task model**.

Target Data Set	**Pearson's r**
EmpCon	0.48
OnSup	0.29
Med	0.28
OffChe	0.14
Avg	0.30

Table 18: Prediction results (Pearson's r) for nonlinear multi-task models based on LIWC embeddings. The first column is the target data set for the respective model that was trained on the remaining three data sets.

Finally, we included the results for the **LIWC-based linear multi-task model** in Table 19.

Target Data Set	**Pearson's r**
EmpCon	0.31
OnSup	0.45
Med	0.26
OffChe	0.06
Avg	0.27

Table 19: Prediction results (Pearson's r) for linear multi-task models based on LIWC embeddings. The first column is the target data set for the respective model that was trained on the remaining three data sets.