A Semantics-based Approach to Disclosure Classification in User-Generated Online Content

Chandan Akiti, Anna Squicciarini, Sarah Rajtmajer
Pennylvania State University
{cra5302, acs20, smr48}@psu.edu

Abstract

As users engage in public discourse, the rate of voluntarily disclosed personal information has seen a steep increase. So-called self-disclosure can result in a number of privacy concerns. Users are often unaware of the sheer amount of personal information they share across online forums, commentaries, and social networks, as well as the power of modern AI to synthesize and gain insights from this data. This paper presents an approach to detect emotional and informational self-disclosure in natural language.We hypothesize that identifying frame semantics can meaningfully support this task. Specifically, we use Semantic Role Labeling to identify the lexical units and their semantic roles that signal self-disclosure. Experimental results on Reddit data show the performance gain of our method when compared to standard text classification methods based on BiLSTM, and BERT. In addition to improved performance, our approach provides insights into the drivers of disclosure behaviors.

1 Introduction

With the growth of social networking sites and increased user engagement with public discourse online, heightened rates of user disclosure of personal information (henceforth, self-disclosure) have raised privacy and security concerns. Prior research (Keep et al., 2012) suggests that self-disclosure may be more common online thanks to the relative anonymity afforded in this environment and the lack of non-verbal cues to signal thoughts or feelings. Users are often unaware of the aggregate amount of personal information they share, as well as the power of modern AI to synthesize and gain insights from this data.

Automating the process of identification and classification of private information in text is challenging (Abril et al., 2011). A large volume of textual data needs to be processed, and a number of real-time requirements need to be met (Agerri et al., 2015), (Singh and Nene, 2013), and significant ambiguities arise from nuanced use of natural language.

Figure 1: SRL of a sentence containing emotional disclosure.

In this work, we adopt the existing framework of Semantic Role Labeling to support self-disclosure identification and classification. Semantic role labeling (SRL) is a process which aims to recognize all predicate-argument pairs along with their roles in a given sentence and its predicates (usually verbs). SRL is a task with numerous applications to Natural Language Processing (NLP) like Question-Answering (Abujabal et al., 2017), Information Extraction (Christensen et al., 2011), Machine Translation (Xiong et al., 2012), text-to-scene generation (Coyne et al., 2012), dialog systems (Chen et al., 2013) and social-network extraction (Agarwal et al., 2014). We hypothesize that the inclusion of semantic frames can provide valuable context for the detection of self-disclosure. Our code is available here[1].

Self-disclosure in social media can take two non-exclusive forms: emotional disclosure, in which the user reveals their feelings towards something or someone; and informational disclosure, where the user reveals objective personal information, e.g., age, career, or address. Following, we propose an approach detecting emotional and informational self-disclosure in text. Specifically, we leverage the structured representations of frame semantics. Our

[1] https://github.com/chandan047/SemanticDisclosure

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3490–3499

method outperforms standard classification methods based on CNN, BiLSTM, and BERT by 9% for emotional disclosure and 2% for informational disclosure.

2 Related Work

Detection of private and sensitive information from user texts has been studied extensively. However, approaches to date appear to be either confined to specific application domains or targeted to specific identifying attributes. Many automated methods for detection of self-disclosure rely on the presence of first-person pronouns, disregard context, and suffer from poor generalizability (Caliskan Islam et al., 2014; Wang et al., 2016a; Vasalou et al., 2011; Bak et al., 2014; Chow et al., 2008; Choi et al., 2013).

Fundamentally, most studies equate disclosure to the revelation of explicitly private information (Wang et al., 2016b). We posit that this frame is insufficient to capture the breadth of victimization that can result from voluntarily shared personal information (e.g., cyberbullying (Joinson and Paine, 2007)), and critically, harms supported by increasingly powerful inference algorithms operating on massive-scale longitudinal datasets (e.g., targeting, manipulation (Paramarta et al., 2018)).

Recent advances in language models have shown improved applicability to classification tasks. Vaswani et al. (2017) introduced a deep bidirectional transformer (BERT) which provided state-of-the-art results on numerous NLP tasks (Devlin et al., 2018). We use BERT as a baseline in this paper. Mehdy et al. (2019) proposed a method to detect disclosures of private information in natural language text through linguistically-motivated artificial neural networks. However, these models do not provide insights into the drivers of disclosure. Sundar et al. (2020) propose heuristics to predict information disclosure, but these heuristics are not exhaustive.

Gildea and Jurafsky (2002) first introduced the task of detecting the semantic frames evoked in text (Semantic Role Labeling; SRL), along with their arguments, formalized in Baker et al. (2007). There are several SRL annotation conventions, such as PropBank (Palmer et al., 2005) and FrameNet (Baker et al., 2007). Propbank provides a more general role labeling, whereas FrameNet provides much denser annotations with more than 1200 frame types. Several studies have explored (Guan et al., 2019) SRL with deep learning techniques. Sikos and Padó (2018) shows that the semantic frames defined in FrameNet can be extended across languages.

Apart from this, several studies have applied SRL features to other Natural Language Processing tasks. Marzinotto et al. (2019) adapted a FrameNet semantic parser for spoken language understanding using adversarial learning. Abujabal et al. (2017) used semantic parsing to generate templates for question answering tasks. Christensen et al. (2011) used semantic role labeling to extract relations in the text without predefining domain or vocabulary. Xiong et al. (2012) utilized the predicate-argument structure of semantic role labeling to enhance Machine Translation. Coyne et al. (2012) extends the existing FrameNet database to bridge visual cues with semantic frames for the text-to-scene generation task. Chen et al. (2013) used semantic parsers to induce and fill semantic slots in dialog systems automatically. While, Agarwal et al. (2014) extract social networks from unstructured text using the FrameNet-defined tree kernel representations.

Our work is motivated in part by Tenney et al. (2019). Authors show that BERT contains elements of the natural language processing pipeline: POS tagging, parsing, NER, semantic roles, and coreference. We explore semantic role labeling specifically for the disclosure detection problem.

3 Frame Semantics

The theory of Frame Semantics asserts that people understand the meaning of words largely by the *frames* which they evoke. The frames represent story fragments, which serve to connect a group of words to a bundle of meanings; for example, the term *avenger* evokes the *Revenge* frame, which describes a complex series of events and a group of participants. The study of Frame Semantics attempts to define frames and the "participants and props" involved in each of them.

A frame is composed of lexical units with frame elements. A *lexical unit (LU)* is a pairing of a word with a meaning. Typically, each sense of a word belongs to a different semantic frame, a script-like conceptual structure that describes a particular type of situation, object, or event along with its participants and props. For example, the *Apply_heat* frame describes a common situation involving a *Cook*, *Food*, and a *Heating Instrument*. These semantic roles are referred to as *frame elements (FEs)*.

Frame-evoking words are LUs in the *Apply_heat* frame. This frame is evoked by words such as bake, blanch, boil, broil, brown, simmer, steam, etc.

The FrameNet (Baker and Sato, 2003; Ruppenhofer et al., 2006) lexical database currently contains more than 13,000 lexical units, around 7,000 of which are hierarchically annotated. A total of approximately 1200 semantic frames are exemplified in more than 200,000 annotated sentences.

4 Frame Semantics for Disclosure Detection

We approach the problem of disclosure detection through the learning of semantic-role based labels common to disclosure. The intuition behind Semantic Role Labeling is to assign *semantic roles* consistent with the frame semantics that are predefined in FrameNet (Baker and Sato, 2003; Ruppenhofer et al., 2006) database. Accordingly, SRL models recover the latent predicate-argument structure of a sentence.

Exemplar sentences and frame semantics are shown in Figures 1 and 2. Target words and text spans are highlighted in the sentence, and their lexical units are shown italicized below. Frames are shown in colored blocks, and frame element segments appear horizontally alongside the frame.

The SRL-labeled sentence in Figure 1, provides an example of a sentence containing emotional disclosure. The frame *Emotion_Active* is invoked by the predicate *"worried"*. This frame has two lexical units containing words *"I"* and *"worried"*. The lexical unit *"I"* is assigned a semantic role of *Experiencer*. We call *Experiencer* a frame-element of *Emotion_Active* frame. Clearly, *Emotion_Active* with an *Experiencer* as *"I"* leads to a self-disclosure of emotion.

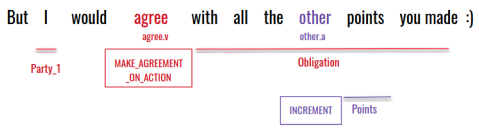

Figure 2: SRL of a sentence containing informational disclosure

Figure 2 shows the case of multiple frames invoked by different predicates. The *Make_Agreement_On_Action* is invoked by the predicate *"agree"*. This frame has multiple lexical units, but two frame elements. The frame element *Party_1* is assigned to *"I"* and *Obligation* is assigned to the span *"with all the other points you made"*. This frame support informational self-disclosure.

Our model predicts disclosure in a sentence based on the semantic frames present. We formulate our disclosure classification model as follows. A sentence S contains a set of semantic frames $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_m\}$ where $m \leq M$. Every semantic frame $\mathcal{F}_j$ has a frame identification I_j and frame elements set $E_j = \{E_{j1}, E_{j2}, \ldots, E_{jk}\}$ where $k \leq K$ and $E_{jk} \in \mathcal{E}$ represents k^{th} frame element of j^{th} semantic frame in the sentence. $\mathcal{E}$ is a set of pre-defined frame elements in FrameNet. In our formulation of the problem, the sentence S contains a disclosure if at least one of the frames $\mathcal{F}$ contained in S is associated with disclosure, according to our classifier. Formally,

$$D(S) = \sigma\left(\max_{f \in \mathcal{F}} D'(f)\right) \quad (1)$$

where $\mathcal{F}$ is the set of semantic frames in the sentence. D' is a disclosure-frame classification function which takes frame $f \in \mathcal{F}$ as the argument. D is the disclosure-sentence classification function for S and σ is the classification function.

5 Semantic Frame Embedding

The FrameNet project (Baker and Sato, 2003; Ruppenhofer et al., 2006) has developed a lexicon of more than 1,200 semantic frames, and a corpus of sentences annotated with frames. We use the FrameNet database to extract semantic frames from the sentences in our dataset. Frame-semantic parsing is a pipeline of three sub-tasks: predicate identification (Which words evoke the frames?); frame identification (Which frames does each predicate evoke?); and argument (frame-element) identification (Which span of the text provides possible roles from $\mathcal{E}$?). Target identification is usually a classification problem.

For the purpose of frame semantics extraction, we use open-SESAME (SEmi-markov Softmax-margin ArguMEnt parser; Swayamdipta et al. (2017)), a framework that provides a pipeline for the three steps mentioned above. Open-SESAME uses Bi-LSTM to classify whether each word in the sentence is a predicate. For each detected predicate (mapped to all possible spans in the sentence), the framework classifies the semantic frame invoked using another Bi-LSTM. Then the framework uses segmental RNN (SegRNN; Kong et al. (2015)) for

predicting frame-elements for the semantic frames detected in the previous step.

5.1 Frame-semantic feature representation

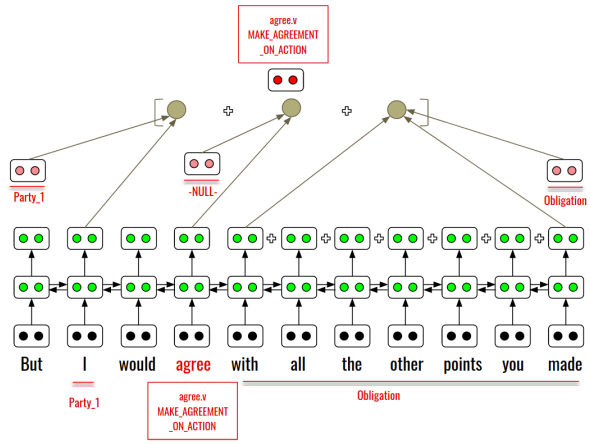

Figure 3: Frame embedding with extracted SRL from the open-SESAME (Swayamdipta et al., 2017) pipeline. The input token embeddings are shown as black, and the input frame and frame-element embeddings are shown in purple. The token bi-LSTM hidden states are shown in green. Grey cells represent the frame-element embedding. Finally, the embedding of the frame is the sum of all frame-elements concatenated (in the figure) with the embedding of frame identification name.

A semantic role labeling of a sentence S is set of semantic frames $\mathcal{F} = \{\mathcal{F}_1, \ldots, \mathcal{F}_m\}$ where $m \leq M$. Every semantic frame $\mathcal{F}_j$ has a name I_j and frame elements set $E_j = \{E_{j1}, \ldots, E_{jk}\}$ where $k \leq K$. E_{jk} is the k^{th} frame element of j^{th} semantic frame in the sentence S. A frame-element has a name and a span of the sentence.

We represent the frame semantics in sentence S as the set of embeddings for each semantic frame. A semantic frame is represented as a combination of two parts. The first part is the **predicate embedding** P_j, a concatenation of the word embedding for predicate w_j and the word embedding for frame name I_j.

$$P_j = [I_j; w_j] \tag{2}$$

where I_j is the frame name, w_j is the corresponding predicate.

The second part of the semantic frame embedding is **frame-element embedding** FE_j. The embedding for frame-elements set E_j is calculated as the combination of embeddings for all frame elements $E_{j1}, E_{j2}, \ldots, E_{jK}$. The embedding for each frame element E_{jk} is a concatenation of word embeddings for frame-element name and corresponding span.

$$FE_j = \frac{1}{K} \sum_{k=1}^{K} [E_{jk}; s_{jk}] \tag{3}$$

where s_{jk} is the span for k^{th} frame-element and E_{jk} is the frame-element name.

Thus, a frame is embedded as

$$\mathcal{F}_j = W_f[P_j; FE_j] + b_f \tag{4}$$

where P_j is the predicate-frame embedding and FE_j is the frame-elements embedding. W_f, b_f are weight and bias parameters for a fully connected layer.

Thus, the sentence S with M frames has a frame semantic representation as

$$S = [\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_M]. \tag{5}$$

5.2 Classification model

In this section, we discuss our model for the disclosure function described in Eq.1. The semantic frame representations extracted in Section 5.1 are stacked to form the sentence representation $[\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_M]$. We model the function D' as a multi-layer perceptron that is applied on each semantic frame. The max function in Eq.1 is modeled as the *MaxPool* layer that outputs maximum activation from all frames.

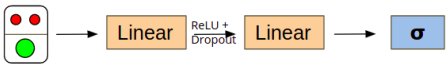

Figure 4: Classification model: The frame representation is shown with two red cells. The green cell is the sentence representation. σ is SoftMax layer whose output is sent to the max function.

The output of max layer is normalized with SoftMax again. The maximum likelihood loss of the final two outputs is optimized.

Through this approach, we classify *emotional disclosure* and *informational disclosure*. Gold training data from Affcon 2020 is used as the labeled data.

6 Experimental Evaluation

6.1 Dataset

Reddit[2] is a popular discussion forum platform consisting of a large number of subreddits focus-

[2] reddit.com

ing on different topics and interests. The Reddit dataset (Jaidka et al., 2020) consists of 12,860 labeled sentences and 5,000 unlabeled sentences. The sentences are sampled from comments in two subreddits: r/CasualConversation, a 'friendlier' sub-community where people are encouraged to share what is on their mind about any topic, and r/OffMyChest, intended as a mutually supportive community where *'deeply emotional things people cannot tell others they know can be told'*. The topics of the collected posts are limited to relationships, with the following tags: "wife"; "girlfriend"; "gf"; "husband"; "boyfriend" and "bf". The statistics of the data from each community are detailed in Table 1.

Label	r/OffMyChest	r/CasualConversation
Emo	2449	1499
Info	2749	1742
Total	**7613**	**5247**

Table 1: Dataset statistics: This table shows the number of Emotional and Information disclosure sentences.

The dataset contains six gold labels for each sentence: *emotional disclosure; information disclosure; support; general support; information support; and, emotional support*. For the purpose of this paper, we only use gold labels of emotional and information disclosure.

Label	Frequency
Emotional	0.31
Informational	0.38

Table 2: Dataset statistics: label frequency.

The Open-SESAME framework assumes a grammatically correct sentence input for which parts-of-speech can be extracted easily. However, Reddit data is prone to ungrammatical sentences, particularly in long paragraphs. To ameliorate this, we exclude from our analysis sentences with more than 50 words. For our dataset, our model provides frame-semantics with $M = 6$ and $K = 5$.

6.2 Semantic frames and frame elements closely linked with SD

In this section, we study frame relevance to emotional and informational disclosure. We operationalize the relevance of a frame as a correlation with emotional (or informational) disclosure. For this analysis, we take the Term Frequency - Inverse Document Frequency (TF-IDF) representation of each sentence with respect to the semantic frames evoked in the sentences. With the TF-IDF representation as features of the sentences, we calculate and normalize feature importance using a random forest classifier. In Figure 5, we show the semantic frame relevance for the top 40 most relevant semantic frames to emotional and informational disclosure.

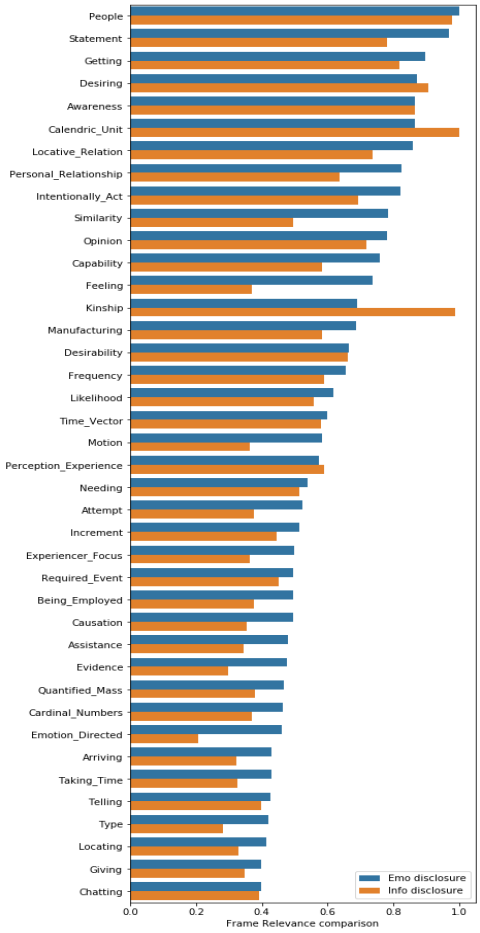

Figure 5: Frame relevance with Emotional disclosure and Informational disclosure

As reported, frames related to emotion (*Feeling*, *Emotion directed*) show high correlation with emotional disclosure as compared to informational disclosure. For example, *"The parent is INFURIATED"* is a sentence containing the word 'INFURIATED' which evokes 'Emotion directed' frame.

We also see the frame 'Kinship,' a frame that is evoked by kinship relational words has a high correlation to the informational disclosure as compared to emotional disclosure. For example, *"Matilde is Gilberto's COUSIN"* is a sentence containing the word 'COUSIN' which evokes 'Kinship' frame. The information disclosed in this sentence is Matilde and Gilberto are related by a kinship relation called cousin.

Although the disclosure is not directly correlated with frame invoked, the observations above are strong motivation for us to explore our method in Section 5 to embed this information in our classification model for high performance.

7 Baseline

7.1 B1: biLSTM model

We use the text classification model Bidirectional LSTM as our baseline. This model is an extension of traditional LSTM. Bidirectional LSTM trains two (instead of one) LSTM on the input sequence. The first input sequence is as-is and the second is on a reversed copy of the input sequence. This provides a forward and backward context for each token in the input sentence. *biLSTM*s provide competitive performance on text classification tasks. (Devlin et al., 2018)

We use *Glove.6B.200d* embedding for the input tokens. The model is trained with an Adam optimizer and a learning rate of 5e-4 for ten epochs. The results are as shown in Tables 3 and 4. As shown, informational disclosure labels are detected with higher accuracy, and achieve 0.61 F1 score, mostly due to a higher precision rate than what we report for emotional disclosure labels.

7.2 B2: BERT model

BERT is a deeply bidirectional, unsupervised language representation,pre-trained (Devlin et al., 2018) using only a plain text corpus from BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words). This makes it particularly suitable for our baseline task, as it allows us to input training text as-is, without imposing predefined and possibly biased features or setting hyper-parameters that would require further analysis.

We train the *bert-base-uncased* version of BERT (12 layers, with a hidden size of 768 and 12 self-attention heads) with an Adam optimizer and learning rate $1e-5$ for two epochs. The results of this baseline method are shown in Tables 3 and 4.

8 Model Classification Results

In our model **A1**, we use *Glove.6B.200d* embedding for the input tokens. The frame-semantic feature representation of each frame in the sentence is extracted, as described in section 5.1. A regular dropout layer will not help regularize the activations independently for the equivalent features in different frames. Accordingly, we add a spatial 1D-Dropout layer after the frame-semantic embedding layer to help regularize the model. We apply 1D convolution with 32 kernels after the Dropout layers. The output of this layer is passed to a MaxPool layer applied to each kernel output. Finally, the output of the MaxPool layer is passed to a fully-connected layer and a classification (Sigmoid) layer.

Another variation of our model comes with replacing the *Glove + biLSTM* layers with an *ELMO* (Peters et al., 2018) or *BERT* (Devlin et al., 2018) embedding. ELMO is a shallow bidirectional model. BERT, unlike ELMO, is a deeply bidirectional model. We present two results with these layers as the contextual word embedding layers for extracting frame-semantic representations.

Model	Precision	Recall	F1-score
B1	0.43	0.67	0.53
B2	0.48	0.68	0.57
A1	0.55*	**0.72**	0.62*
A1+ELMO	0.52	0.69	0.59
A1+BERT	**0.57**	0.71*	**0.63**

Table 3: 10-fold cross-validation results for emotion disclosure.

Model	Precision	Recall	F1-score
B1	0.56	0.67	0.61
B2	**0.60**	0.64	0.62
A1	0.57	0.69*	0.63*
A1+ELMO	0.58	0.66	0.62
A1+BERT	0.59*	**0.69**	**0.64**

Table 4: 10-fold cross-validation results for informational disclosure.

For the BERT version of our model, we take frame identification name embedding as the hidden state of the BERT last layer when the input is a tokenized version of the name. The results in Tables 3 and 4 indicate the performance of our model with BERT embeddings. There is a marginal improvement in the F1-score compared to the model with Glove embedding.

9 Ablation

We study the effect of frame-semantic features on the classification task. Our model, when compared with the biLSTM model, improves the F1-score

on Emotional Disclosure by **9%** and Informational Disclosure by **2%**. We considered the contextual word embeddings of tokens in the sentence, frame identification name, and frame elements. This ablation study is to measure the effect of frame-elements on the classification task qualitatively.

Label	Model	Precision	Recall	F1
Emo	B1	0.43	0.67*	0.53
Emo	A1-FE	0.54*	0.66	0.60*
Emo	A1	**0.55**	**0.72**	**0.62**
Info	B1	0.56	0.67*	0.61*
Info	A1-FE	0.57*	0.65	0.61
Info	A1	**0.57**	**0.69**	**0.63**

Table 5: Ablation study for model without frame elements in frame-semantic feature representations. Comparing results of three models B1, A1-FE, and A1 with average scores on 10-fold cross validation.

We use a smaller version **A1-FE** of our model. In this version, we remove the frame-element embeddings (from gray cells in Figure 3) from the semantic frame feature representation. Thus, frame-semantic representation concatenates the embedding of a frame identification name and the predicate. The classification model remains the same as **A1** (Figure 4).

Our results are summarized in Table 5. Emotional disclosure classification improves the F1-score by **7%** with frame identification features. This improves further (**2%**) with frame-element features. This result indicates that frame identification names carry maximum information pertained to Emotional Disclosure.

Informational disclosure classification performs similarly with or without frame-identification features. However, there is an improvement in F1-score by **2%** with frame-element features. This improvement suggests that we cannot infer informational disclosure from the frame identification name alone. Frame-elements are crucial for detecting informational disclosure.

10 The Role of Conversation in Self-Disclosure

In this section, we explore the effect of peer influence features on disclosure detection in the conversation. Given the highly contextual and inherently *social* character of self-disclosure, we are motivated to explore peer influence as a meaningful signal for this behavior. We incorporate peer effects in models aiming to detect and predict disclosure in conversation. We touch upon some early findings in this direction. Practically we develop a model that addresses the problem of predicting disclosure in a given comment using the comment text and peer influence features extracted from previous comments.

We have completed an early exploration of conversational modeling of the effects of peer influence in Reddit conversations. We augmented our original Reddit dataset (Jaidka et al., 2020) with the missing responses from the original comments using Python Reddit API Wrapper (PRAW) [3]. We sample 1200 conversations (about 1600 users) from the comment trees and manually annotate the emotional and informational disclosure in comments using Mechanical Turk with consensus from three workers. We labeled the dataset into three labels: *No Disclosure* vs *Low Disclosure* vs *High Disclosure* for each of emotional and informational disclosure. We then calculate reliability metrics [4] which indicate high reliability scores for binary: *No Disclosure (no)* vs *Disclosure (low or high)* labels.

Type	Reliability metric	Info	Emo
No vs Low vs High	Fleiss Kappa	0.484	0.242
	Gwet's AC1/AC2	0.631	0.317
Binary	Fleiss Kappa	**0.653**	**0.394**
	Gwet's AC1/AC2	**0.701**	**0.644**

Table 6: Table indicating high reliability scores for binary labels of emotional and informational disclosure

A Reddit post is composed of a *post text* written by an *author*, *comments* and *votes*. A comment is composed of *comment text* written by an *author*, *reply comments* and *votes*. Recursive comments within each post form a comment tree. We sample conversations from this comment tree by recursively taking the comments where the successor comment is a direct reply to the predecessor comment. Unlike the original dataset (Jaidka et al., 2020) which provides annotations for sentences sampled from the comments, here we classify the entire comment.

[3] https://praw.readthedocs.io/en/latest/

[4] https://www.ncbi.nlm.nih.gov/books/NBK92295/table/methods.t2/

Disclosure	Positive	Total
Emo Disclosure	826	1200
Info Disclosure	543	1200

Table 7: Label statistics for emotional and informational disclosure

Our model is a simple modification of *BertForSequenceClassification* from Facebook's HuggingFace library (Wolf et al., 2019). The model predicts disclosure in the fifth comment using the comment text and the following peer influence factors: number of unique self-disclosures in the last four comments; whether there is a disclosure from the same user in last four comments; the total number of users in the given conversation; previous disclosure; time elapsed since the most recent prior disclosure; and time elapsed since the most recent prior comment. Elapsed time is normalized to [0, 1], where 1 represents one day and any more than a day.

The *bert-base-uncased* architecture of the BERT model is enhanced with the peer influence features listed above. The pooled output of the BERT model of dimension 768 is passed through a dropout layer with a dropout rate of 0.05 and then passed through a linear layer with an output dimension of 16. The peer influence features are appended to this output and passed though a binary classification layer. We train the model using an Adam optimizer with a learning rate of 4×10^{-5} for one epoch.

The emotional disclosure classification model achieves **87.9%** F1-score (representing a **2.2%** improvement on the model without peer influence features) and **0.54** Matthews Correlation Coefficient (MCC) with 5-fold cross validation. Similarly, the informational disclosure classification model achieves **73.3%** F1-score (**1.5%** improvement) and **0.61** MCC with 5-fold cross validation.

These early promising results in affect analysis of peer-influence on disclosure, point to further exploration of frame semantics in conversation/dialogue systems as a promising avenue for future work.

11 Conclusion

In this paper, we have presented a study using semantic role labels to support the detection of voluntarily disclosed private information in a user-generated text. To the best of our knowledge, ours is the first study performing in-depth semantic analysis to facilitate detection and analysis of self-disclosure. In doing so, we have simultaneously improved upon state-of-the-art performance for detection of disclosure in sentences and furnished meaningful semantic information about tagged disclosures. The success of frame semantics in helping to identify sentences containing disclosure is perhaps unsurprising given its power in distilling meaning from groups of individual words. Yet, our models have potential for more insightful analysis, beyond what is presented here. For example, semantic frames across sentences in the comment can be linked in a graph-like structure if the same entities evoke the semantic frames. Moreover, the same can be applied across comments. We will explore these graph-based approaches in future work.

References

Daniel Abril, Guillermo Navarro-Arribas, and Vicenç Torra. 2011. On the declassification of confidential documents. In *Proceedings of the 8th International Conference on Modeling Decisions for Artificial Intelligence*, MDAI'11, page 235–246, Berlin, Heidelberg. Springer-Verlag.

Abdalghani Abujabal, Mohamed Yahya, Mirek Riedewald, and Gerhard Weikum. 2017. Automated template generation for question answering over knowledge graphs. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1191–1200, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Apoorv Agarwal, Sriramkumar Balasubramanian, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2014. Frame semantic tree kernels for social network extraction from text. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 211–219, Gothenburg, Sweden. Association for Computational Linguistics.

Rodrigo Agerri, Xabier Artola, Zuhaitz Beloki, German Rigau, and Aitor Soroa. 2015. Big data for natural language processing: A streaming approach. *Knowledge-Based Systems*, 79:36 – 42.

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, Doha, Qatar. Association for Computational Linguistics.

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International*

Workshop on Semantic Evaluations (SemEval-2007), pages 99–104, Prague, Czech Republic. Association for Computational Linguistics.

Collin F. Baker and Hiroaki Sato. 2003. The FrameNet data and software. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, page 161–164.

Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. 2014. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, WPES '14, pages 35–46, New York, NY, USA. ACM.

Y. Chen, W. Y. Wang, and A. I. Rudnicky. 2013. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 120–125.

Dongjin Choi, Jeongin Kim, Xeufeng Piao, and Pankoo Kim. 2013. Text analysis for monitoring personal information leakage on twitter. *Journal of Universal Computer Science*, 19(16):2472–2485.

Richard Chow, Philippe Golle, and Jessica Staddon. 2008. Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 893–901, New York, NY, USA. ACM.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the Sixth International Conference on Knowledge Capture*, K-CAP '11, page 113–120, New York, NY, USA. Association for Computing Machinery.

Bob Coyne, Alex Klapheke, Masoud Rouhizadeh, Richard Sproat, and Daniel Bauer. 2012. Annotation tools and knowledge representation for a text-to-scene system. In *Proceedings of COLING 2012*, pages 679–694, Mumbai, India. The COLING 2012 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288.

Chaoyu Guan, Yuhao Cheng, and Hai Zhao. 2019. Semantic role labeling with associated memory network. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3361–3371, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

AN Joinson and CB Paine. 2007. Self-disclosure, privacy and the internet. oxford handbook of internet psychology.

Melanie Keep, Yu Sun Bin, and Andrew Campbell. 2012. Comparing online and offline self-disclosure: A systematic review. *Cyberpsychology, behavior and social networking*, 15:103–11.

Lingpeng Kong, Chris Dyer, and Noah A. Smith. 2015. Segmental recurrent neural networks.

Gabriel Marzinotto, Geraldine Damnati, and Frédéric Béchet. 2019. Adapting a FrameNet Semantic Parser for Spoken Language Understanding Using Adversarial Learning. In *Interspeech 2019*, pages 799–803, Graz, Austria. ISCA.

Nuhil Mehdy, Casey Kennington, and Hoda Mehrpouyan. 2019. Privacy disclosures detection in natural-language text through linguistically-motivated artificial neural networks.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Valentinus Paramarta, Muhammad Jihad, Ardhian Handoyo, Ika Hapsari, Puspa Sandhyaduhita, and Achmad Hidayanto. 2018. Impact of user awareness, trust, and privacy concerns on sharing personal information on social media: Facebook, twitter, and instagram. pages 271–276.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2006. Framenet ii: Extended theory and practice.

Jennifer Sikos and Sebastian Padó. 2018. Using embeddings to compare FrameNet frames across languages. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 91–101, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jayveer Singh and Manisha J. Nene. 2013. A survey on machine learning techniques for intrusion detection systems.

S. Shyam Sundar, Jinyoung Kim, Mary Beth Rosson, and Maria D. Molina. 2020. Online privacy heuristics that predict information disclosure. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Asimina Vasalou, Alastair J. Gill, Fadhila Mazanderani, Chrysanthi Papoutsi, and Adam Joinson. 2011. Privacy dictionary: A new resource for the automated content analysis of privacy. *Journal of the American Society for Information Science and Technology*, 62(11):2095–2105.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016a. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 74–85, New York, NY, USA. ACM.

Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016b. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work Social Computing*, CSCW '16, page 74–85, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Deyi Xiong, Min Zhang, and Haizhou Li. 2012. Modeling the translation of predicate-argument structure for SMT. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 902–911, Jeju Island, Korea. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.