Self-Reflective Sentiment Analysis

Benjamin Shickel
University of Florida
Gainesville, FL 32611
shickelb@ufl.edu

Martin Heesacker
University of Florida
Gainesville, FL 32611
heesack@ufl.edu

Sherry Benton
TAO Connect, Inc.
Gainesville, FL 32601
sherry.benton@taoconnect.org

Ashkan Ebadi
University of Florida
Gainesville, FL 32611
ashkan.ebadi@ufl.edu

Paul Nickerson
University of Florida
Gainesville, FL 32611
pvnick@ufl.edu

Parisa Rashidi
University of Florida
Gainesville, FL 32611
parisa.rashidi@ufl.edu

Abstract

As self-directed online anxiety treatment and e-mental health programs become more prevalent and begin to rapidly scale to a large number of users, the need to develop automated techniques for monitoring patient progress and detecting early warning signs is at an all-time high. While current online therapy systems work based on explicit quantitative feedback from various survey measures, little attention has been paid thus far to the large amount of unstructured free text present in the monitoring logs and journals submitted by patients as part of the treatment process. In this paper, we automatically categorize patients' internal sentiment and emotions using machine learning classifiers based on n-grams, syntactic patterns, sentiment lexicon features, and distributed word embeddings. We report classification metrics on a novel mental health dataset.

1 Introduction

As mental health awareness becomes more widespread, especially among at-risk populations such as young adults and college-aged students, many institutions and universities are beginning to offer online anxiety and depression treatment programs to supplement traditional therapy services. A key component of these largely self-directed programs is the regular completion of journals, in which patients describe how they are feeling. These journals contain a wide variety of information, including a patient's specific fears, worries, triggers, reactions, or simply status updates on their emotional state. At current time, these journals are either reviewed by therapists (who are vastly outnumbered by the users) or left unused, with the assumption that simply talking about negative emotions is therapy in and of itself. We see a large and novel opportunity for applying natural language techniques to these unstructured mental health records. In this paper, we focus on analyzing the sentiment of patient text.

The largest motivator of existing sentiment analysis research has arguably been the detection of user sentiment towards entities, such as products, companies, or people. We define this type of problem as *external* sentiment analysis. In contrast, when working in the mental health domain (particularly with self-reflective textual journals), we are trying to gauge a patient's *internal* sentiment towards their own thoughts, feelings, and emotions. The differences in goals, types of sentiment, and distribution of polarity presents unique challenges for applying sentiment analysis to this new domain.

One key aspect that sets our task apart from traditional sentiment analysis is our treatment of polarity classes. Traditionally, sentiment is categorized as either *positive*, *negative*, or *neutral*. In contrast, we subdivide the *neutral* polarity class into two distinct classes: *both positive and negative* and *neither positive nor negative*. We justify this decision based on several studies showing the independent dimensions of positive and negative affect in human emotion (Warr et al., 1983; Watson et al., 1988; Diener et al., 1985; Bradburn, 1969), and feel that is a more appropriate framework for our domain. This choice represents a novel characterization of sentiment analysis in mental health, and is one we hope

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 23–32,
San Diego, California, June 16, 2016. ©2016 Association for Computational Linguistics

to see made in future studies in this domain.

Our primary focus in this paper is on the automatic and reliable categorization of patient responses as *positive*, *negative*, *both positive and negative*, or *neither positive nor negative*. Such a system has far-reaching implications for the online therapy setting, in which automatic language analysis can be incorporated into existing patient evaluation and progress monitoring, or serve as an early warning indicator for patients with severe cases of depression and/or risk of suicide. Additionally, tools based on this type of internal sentiment analysis can provide immediate feedback on mental health and thought processes, which can become distorted and unclear in patients stuck in anxiety or depression. In the future, sentiment-based mental health models can be incorporated into the characterization and treatment of patients with autism, dementia, or other broadly-defined language disorders.

In short, our main contributions are summarized by the following:

- We present a novel sentiment analysis dataset, annotated by psychology experts, specifically targeted towards the mental health domain.
- We introduce the notion of subdividing the traditional *neutral* polarity class into both a dual polarity sentiment (*both positive and negative*) and a *neither positive nor negative* sentiment.
- We identify the unique challenges faced when applying existing sentiment analysis techniques to mental health.
- We present an automatic model for classifying the polarity of patient text, and compare our work to models trained on existing sentiment corpora.

2 Related Work

From a technical point of view, our methods fall squarely in the realm of sentiment analysis, a field of computer science and computational linguistics primarily concerned with analyzing people's opinions, sentiments, attitudes, and emotions from written language (Liu, 2010). In our paper, we apply sentiment analysis and polarity detection techniques to the largely untapped mental health domain.

In the past decade, sentiment analysis techniques have been applied to a wide variety of areas. Although the majority of work has dealt in areas outside of mental health, we must discuss the bulk of previous sentiment analysis research, from which our techniques are derived.

Given the explosive rise in popularity of social media platforms, a large number of studies have focused on user sentiment in microblogs such as Twitter (Barbosa and Feng, 2010; Pak and Paroubek, 2010; Agarwal et al., 2011; Kouloumpis et al., 2011; Nielsen, 2011; Wang et al., 2011; Zhang et al., 2011; Montejo-Ráez et al., 2012; Spencer and Uchyigit, 2012; Montejo-Ráez et al., 2014; Tang et al., 2014). Other studies have explored user sentiment in web forum opinions (Abbasi et al., 2008), movie reviews (Agrawal and Siddiqui, 2009), blogs (Melville et al., 2009), and Yahoo! Answers (Kucuktunc et al., 2012). As we will show, the models proposed in all of these works cannot be directly transferred to polarity detection in mental health (as sentiment analysis remains a largely domain-specific task), but our initial techniques are largely based on these previous works.

Although the majority of sentiment analysis has focused on user opinions towards entities, there are studies in domains more directly related to our area. One such study analyzed the sentiment of suicide notes (Pestian et al., 2012). Another mined user sentiment in MOOC discussion forums (Wen et al., 2014).

Sentiment analysis and polarity detection techniques are widely varied (Mejova and Srinivasan, 2011; Feldman, 2013), and as this research area is still garnering a great deal of interest, many studies have proposed novel methods. These include topic-level sentiment analysis (Nasukawa and Yi, 2003; Kim and Hovy, 2004), phrase-level sentiment analysis (Wilson et al., 2009), linguistic approaches (Wiegand et al., 2010; Benamara et al., 2007; Tan et al., 2011), semantic word vectorization (Maas et al., 2011; Tang et al., 2014), various lexicon-based approaches (Taboada et al., 2011; Baccianella et al., 2010), information-theoretic techniques (Lin et al., 2012), and graph-based methods (Montejo-Ráez et al., 2014; Pang and Lee, 2004; Wang et al., 2011). In recent years, approaches based on deep learning architectures have also shown promising results (Glorot et al., 2011; Socher et al., 2013). As a starting point for our new *internal* sentiment analysis framework, in this paper we apply more straightforward approaches based on linear classifiers.

3 Dataset

In this section, we detail the construction of our mental health sentiment dataset. While not yet publicly available, we plan to release our data in the near future.

In order to build a dataset of real patient responses, we partnered with TAO Connect, Inc.[1], an online therapy program designed to treat anxiety, depression, and stress. This program is being implemented in several universities around the country, and as such, the primary demographic is college-aged students.

As part of the TAO program, patients complete several self-contained content modules designed to teach awareness and coping strategies for anxiety, depression, and stress. Additionally, patients regularly submit several types of journals and logs pertaining to monitoring, anxiety, depression, worries, and relaxation. The free text contained in these logs is the source of our dataset. In total, we collected 4021 textual responses from 342 unique patients, with submission dates ranging from April 2014 to November 2015. Patients were de-identified and the collection process was part of an IRB-approved study. Responses typically range from single sentences to a single paragraph, with an average of 39 words per response. We show a complete word count distribution in Figure 1.

To help transform our collection of free text responses into a classification dataset suitable for polarity prediction, we solicited the expertise of three psychology undergraduates (all female) under the supervision of one psychology professor (male) to provide polarity labels for our response documents. The annotators were tasked with reading each individual response, and assigning it a label of *positive*, *negative*, *both positive and negative*, or *neither positive nor negative*. The inter-rater agreement reliability (Cohen's kappa) between annotators 1 and 2 was 0.5, between annotators 2 and 3 was 0.67, and between annotators 1 and 3 was 0.48. The overall agreement reliability between all annotators (Fleiss' kappa) was 0.55. We used a majority-vote scheme to assign a single label to each piece of text, where 62% of the documents had full annotator agreement, 35% had a clear label majority, and only 3% had no majority, in which case we picked the label from the annotator with the best aggregate reliability. Table 1 shows label counts for each annotator, as well as the final count after applying the majority-vote process.

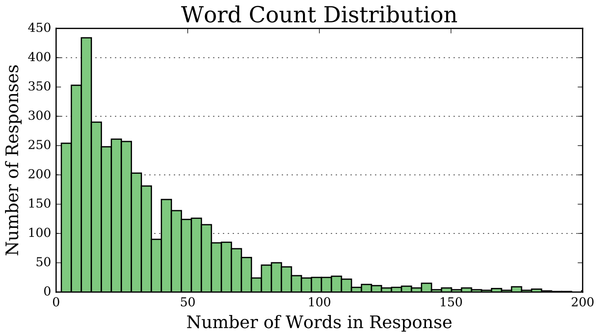

Figure 1: Distribution of word counts per response for our collected dataset. On average, each response contains 39 words, with a minimum of two words and a maximum of 762 words. 30 responses had more than 200 words, which we do not show.

Annotator	POS	NEG	BOTH	NEITHER
Annotator 1	494	2569	556	402
Annotator 2	321	2509	552	638
Annotator 3	531	2152	383	954
Final	414	2545	510	548

Table 1: Label counts per annotator, as well as the the final dataset label counts obtained via a majority-voting scheme. For brevity, we denote the *positive* label as POS, *negative* as NEG, *both positive and negative* as BOTH, and *neither positive nor negative* as NEITHER.

To provide a clearer picture of the types of responses in our dataset, we present one short concrete example of each polarity class below.

- **Positive** - *I tried to say good things for them since I know there was a lot of arguments happening.*
- **Negative** - *I don't do well at parties, I'm not interesting.*
- **Both Positive and Negative** - *I shouldn't have taken things so seriously.*
- **Neither Positive nor Negative** - *I wrote in my*

---

[1] http://www.taoconnect.org/

> *journal, and read till I was tired enough to fall asleep.*

In the above examples, the challenges of applying sentiment analysis and traditional text classification techniques to self-reflective text becomes more apparent. For instance, the *positive* example mentions arguments, typically associated with negative sentiment, while the *negative* example mentions parties, a word usually associated with a positive connotation. Additionally, the *both positive and negative* example exhibits subtle cues that differentiate it from the other three polarity classes.

4 Method

To predict polarity from patient text, we employ several established machine learning and text classification techniques. We begin by preprocessing the annotated patient responses, which we refer to interchangeably as *documents*. We then extract several types of attributes from each response, referred to as *features*. The extracted features and polarity annotations are used to build a logistic regression *classifier*, which is a linear machine learning model we use to predict the final polarity label. In this section, we describe each step in detail.

4.1 Preprocessing

Starting with the raw documents obtained from our data collection process, we apply several traditional preprocessing steps to the text. First, based on experimental feedback, we convert all the text to lowercase and strip all documents of punctuation following a standard tokenization phase. While these are relatively standard steps, it should be explicitly noted that we did *not* remove stop words from our corpus, which is a common preprocessing technique in other domains, due to lowered classification performance. This can be partially explained by the nature of our domain; for example, the phrase “what if” tended to be associated with worrying about the future - traditionally, both of these words are considered stop words and filtered out, losing valuable information for our task.

4.2 Feature Extraction

Next, we extract several types of features from the preprocessed documents. In our experiments, we evaluate classification performance with various feature subsets.

4.2.1 N-Gram Features and POS Tags

As a starting point for our experiments with this new domain, the most numerous of our extracted features are derived from a traditional “bag of n-grams” approach, in which we create document vectors comprised of word unigrams, bigrams, and/or trigram counts. As previous works have shown, this allows the capture of important syntactical information like negation, which would otherwise be missed in a standard “bag of words” (i.e., unigrams only) model.

In order to constrain the scope of later feature subset experiments, we first obtain the n-gram combination resulting in the best performance for our newly created dataset. We denote this optimal n-gram setting as the “n-grams only” model in later experiments. In this experiment, we perform a 10-fold cross-validated randomized parameter search using six possible word n-gram combinations: unigrams, bigrams, trigrams, unigrams + bigrams, bigrams + trigrams, and unigrams + bigrams + trigrams. We split cross-validation folds on responses, as we expect patient responses to be independent over time. All extracted n-gram counts are normalized by tf-idf (term frequency-inverse document frequency), a common technique used for describing how important particular n-grams are to their respective documents. The results of this n-gram comparison experiment are shown in Figure 2, where it is clear that using a combination of unigrams and bigrams resulted in the best performance.

In an effort to capture more subtle patterns of grammatical structure, we also experiment with augmenting each document with each word’s Penn-Treebank part-of-speech (POS) tag. In these experiments, we augment our documents by appending these tags, in order, to the end of every sentence, allowing for our n-gram extraction methods to capture syntactic language patterns. During the tokenization process, we ignore any n-grams consisting of both words and part-of-speech tags.

4.2.2 Sentiment Lexicon Word Counts

One of the more rudimentary sentiment analysis techniques stems from the use of a sentiment dictio-

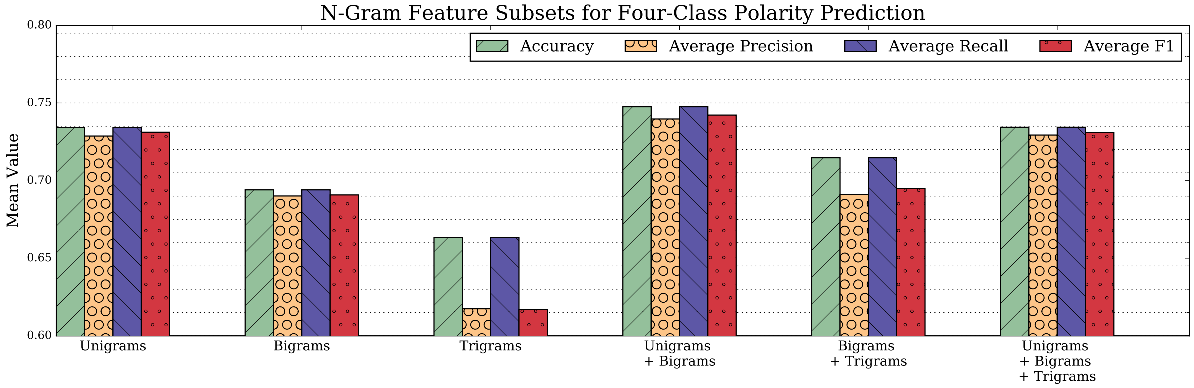

Figure 2: Classification results using only word n-gram features for our 4-class polarity dataset. Results were obtained following a 10-fold cross-validated randomized hyperparameter search. A combination of unigrams and bigrams resulted in the best metrics. As seen by the final cluster, adding trigrams to this subset resulted in a performance decrease. Thus, when we use n-gram features in later experiments, we only consider the combination of unigrams and bigrams.

nary, or lexicon, which is a pre-existing collection of subjective words that are labeled as either *positive* or *negative*. Using the sentiment lexicon from (Liu, 2012)[2], we count the number of positive and negative words occurring in each document and incorporate the counts as two additional features.

4.2.3 Document Word Count

In our initial analysis, we discovered that oftentimes the most negative text responses were associated with a larger word count. Although the correlation is relatively weak across the entire corpus, we nonetheless include a word count of each document as a feature.

4.2.4 Word Embeddings

Based on the recent successes of distributed word representations like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), we sought to harness these learned language models for predicting sentiment polarity. Although primarily used in deep learning architectures, we show that these representations can also be useful with linear models. Unlike our other features, the individual features contained in word embeddings are indecipherable; however, as we show in the results section, they contribute to the overall success of our classification.

In our experiments, we utilize a publicly available Word2Vec model pre-trained on Google News[3], containing 100 billion words. Each unique word in the model is associated with a 300-dimensional vector. For each of our documents, we include the mean word vector derived from each individual word's embedding as 300 additional features.

5 Four-Class Polarity Prediction

Because our new dataset introduces a clear distinction between text labeled as *both positive and negative* and *neither positive nor negative* (traditionally, both of these classes are grouped together as *neutral*), there are no baselines for which to compare our experiments. We offer our results for this scenario as a launching point for future studies on polarity detection in mental health. For this scenario, we show the results of each feature extraction method individually, as well as the results for the combination of all features. All results are evaluated via 10-fold cross-validation, with folds split on responses. Results are shown in Figure 3, where it is clear that optimal performance is achieved using the model trained on all features. Our methods gave rise to an overall classification accuracy of 78%.

From Figure 3, it is apparent that of all individual features, n-grams perform the best. The relatively strong performance of n-gram features tends to align with our expectations, given the widespread use of n-gram features across all types of text classifica-

---

[2] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

[3] https://code.google.com/p/word2vec/

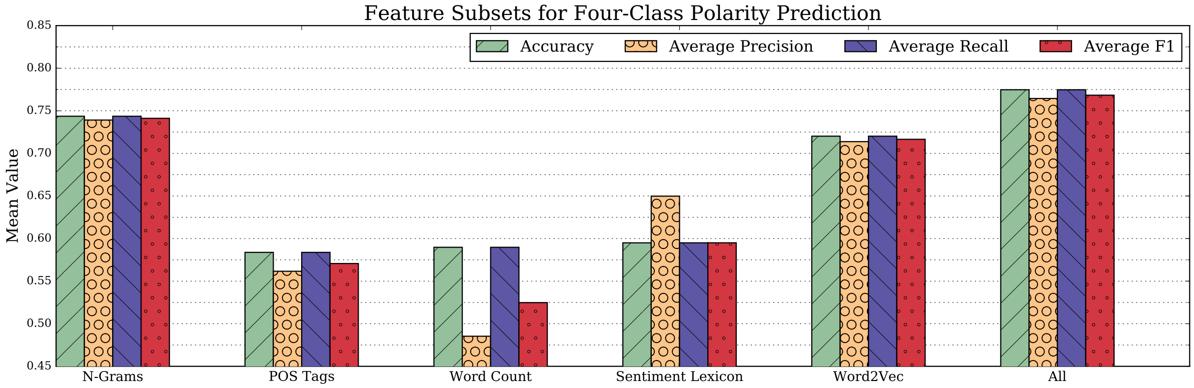

Figure 3: Classification performance for the 4-class polarity prediction task. We show results for each feature set individually, as well as the combination of all features. Using all extracted features results in the highest accuracy, F1, precision, and recall.

tion problems. However, what is more surprising is the relatively weak results for the sentiment lexicon features, given their popularity in modern sentiment analysis. Additionally, the word embedding features also gave rise to better performance than expected, especially considering that we used the Word2Vec embeddings with linear models as opposed to the more traditional deep learning architectures. Finally, we see optimal performance across all metrics when using the combination of all features.

Using the optimal model from Figure 3, we show the individual class metrics for precision, recall, F1, and overall accuracy in Table 2. It is apparent that the *both positive and negative* class proves especially difficult to classify. This is explained in part by the previously mentioned class imbalance issue - when the majority of the corpus is negative, it becomes difficult for the classifier to differentiate between sentiment comprising of *mostly* positive polarity, and sentiment comprising of *some* positive polarity. The low recall of the *both positive and negative* class clearly points towards the need for more research in this area.

Class	Precision	Recall	F1
Positive	0.63	0.32	0.42
Negative	0.74	0.96	0.84
Both	0.58	0.16	0.26
Neither	0.77	0.47	0.59
Overall Accuracy		**0.78**	

Table 2: Polarity prediction results for the full 4-class version of our dataset. For brevity, the polarity class *both positive and negative* is denoted as *Both*, and the class *neither positive nor negative* is denoted as *Neither*.

6 Binary Polarity Prediction

In this section, we experiment with using existing sentiment analysis corpora to perform traditional two-class polarity prediction on our dataset, and compare the results to a cross-validation approach, split on responses, trained on our dataset alone. The primary purpose is to gauge the effectiveness of classifiers trained on existing sentiment corpora as applied to the mental health domain. State of the art sentence-level binary polarity detection accuracy is reported as 85.4% (Socher et al., 2013) using deep learning models and a specialized movie review dataset, and while our models are computationally more simple and use different features, we incorporate such existing corpora in our experiments. Since our full dataset consists of four polarity labels, whereas traditional sentiment analysis only uses two, for these experiments we only consider the responses from our dataset belonging to the *positive* and *negative* classes.

We begin by training our model on existing sentiment datasets only. The first is a large-scale Twitter sentiment analysis dataset[4] which automatically assigns polarity labels based on emoticons present in user tweets (we denote this dataset as “Twitter”). The next is a collection of IMDB movie reviews published by (Maas et al., 2011) at Stanford University[5] (we denote this dataset as “Stanford”). We also use two movie reviews datasets from (Pang et al.,

---

[4] http://help.sentiment140.com/for-students/

[5] http://ai.stanford.edu/ amaas/data/sentiment/

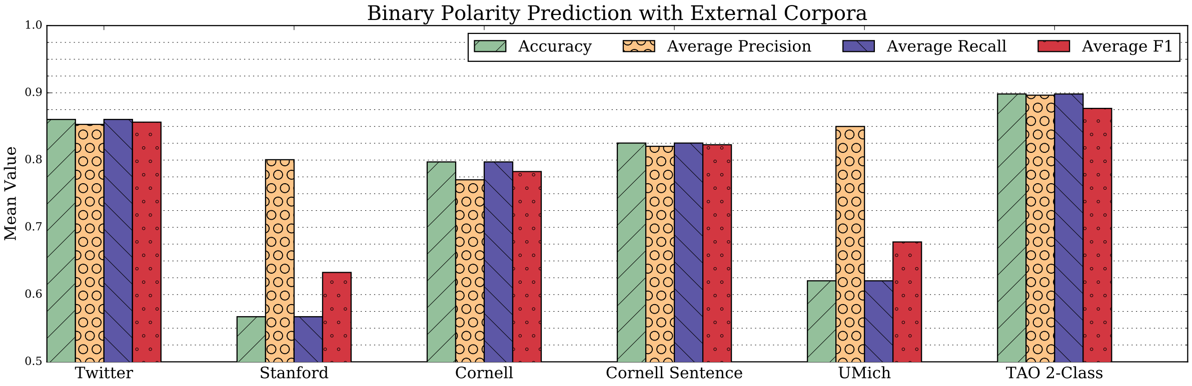

Figure 4: Classification results for the *positive* vs. *negative* prediction setting using 5 external sentiment corpora and cross-validated results on our own binary dataset (TAO 2-Class). The precision, recall, and F1 scores are reported using a weighted average incorporating the support of each class label. For all metrics, training on our dataset (TAO 2-Class) yields better results than using models trained on existing sentiment corpora.

Dataset	# Positive	# Negative
Twitter	797792	798076
Stanford	25000	25000
Cornell	1000	1000
Cornell Sentence	5221	5212
UMich	3995	3091

Table 3: Existing sentiment corpora summary.

2002) at Cornell University[6], where one is geared towards document-level sentiment classification (denoted as "Cornell"), and the other towards sentence-level classification (denoted as "Cornell Sentence"). Our final dataset is a collection of web forum opinions collected by the University of Michigan as part of a Kaggle competition[7] (which we denote as "UMich"). The number of documents of each sentiment class, per dataset, is given in Table 3.

Using all features from the previously outlined extraction process, we train a separate model on each of the five existing sentiment analysis corpora. Optimal hyperparameters for each experiment were selected via a randomized parameter search in conjunction with three-fold cross validation. In each case, the trained models were tested against the binary version of our dataset. Additionally, we perform the same extraction and fine-tuning process to construct a model trained on our new dataset alone. For this experiment, we report the results after a 10-fold cross-validation process split on responses. A summary of accuracy, precision, recall, and F1 score for the binary prediction setting is shown in Figure 4, where it is apparent that the best performance occurs when using our dataset, pointing towards the need for collecting custom mental health datasets for this new type of internal sentiment analysis. Our binary polarity model resulted in 90% classification accuracy.

7 Important Features

In this section, we wish to understand which features are most discriminative in predicting whether a piece of text is *positive*, *negative*, *both positive and negative*, or *neither positive nor negative*. These features (all of which are naturally-interpretable aside from the word embeddings) can serve as useful indicators for therapists and future mental health polarity studies.

To evaluate our features, we examine the weight matrix of a randomized logistic regression classifier trained on our full four-class polarity dataset. The feature weights corresponding to each of the four classes give an idea of the relative importance of each feature, and how discriminative they are as compared to the remaining three classes. We summarize the 10 most important features per class in Table 4.

Much can be gleaned from an informal inspection of these top features. For example, while the words found in the *positive* and *negative* polarity

---

[6] https://www.cs.cornell.edu/people/pabo/movie-review-data/

[7] https://inclass.kaggle.com/c/si650winter11/data

Positive	Negative	Both Positive and Negative	Neither Positive nor Negative
was able	worried	but	work
no anxiety	$RB $VBG	okay	nothing
calm	<W2V-81>	nt worry	$IN $NNP
nothing terrible	$VBN $IN	$NNS $PRP	to the
great	worried about	$VB $RB	slowly
better	worried that	eventually	can
did well	nt do	not as	<W2V-129>
no worries	<W2V-96>	instead	<W2V-230>
not anxious	stressed	although	study
hopeful	<W2V-168>	actually	not sure

Table 4: Top 10 features per class from a randomized logistic regression model, trained on our mental health dataset. Features with a $ symbol are part-of-speech tags (using our POS n-gram method). All individual word embedding features, obtained via a pre-trained Word2Vec embedding, are denoted as <W2V-X>, where X is the dimension index of the embedding vector. The POS tags shown are are as follows: $RB = adverb, $VBG = present participle verb, $VBN = past participle verb, $IN = preposition, $JJ = adjective, $NNS = plural noun, $PRP = personal pronoun, $VB = base form verb, $NNP = singular proper noun.

classes are clearly characteristic of their respective labels (with *negative* words pertaining mostly to worry and stress), the words found in the *both positive and negative* class are more indecisive in nature ('but', 'eventually', 'although', 'actually'). Words from the *neither positive nor negative* class carry less surface-level emotional significance. The part-of-speech patterns are more difficult to interpret, but these results point towards the need for future exploration.

8 Conclusion

In this paper, we introduced the notion of applying sentiment analysis to the mental health domain, and show that existing techniques and corpora cannot be simply transferred to this new setting. We developed baseline classification techniques grounded in the results from previous works, and show the benefit of spending resources on creating new mental health datasets explicitly focused on patient sentiment. We introduced the notion of splitting the polarity class traditionally defined as *neutral* into two sub-classes, and demonstrated the new challenges that decision brings as it pertains to the automatic classification of patient sentiment in mental health text.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, 26(3):1–34.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38.

Shaishav Agrawal and Tanveer J Siddiqui. 2009. Using syntactic and contextual information for sentiment polarity analysis. In *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, pages 620–623.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44.

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V.S. Subrahmanian. 2007. Sentiment analysis: adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 203–206.

Norman M. Bradburn. 1969. The structure of psychological well-being.

E Diener, R J Larsen, S Levine, and R a Emmons. 1985. Intensity and frequency: dimensions underlying positive and negative affect. *Journal of personality and social psychology*, 48(5):1253–1265.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

X Glorot, A Bordes, and Y Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. *Proceedings of the*.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373, Morristown, NJ, USA.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 11)*, pages 538–541.

Onur Kucuktunc, Ingmar Weber, B. Barla Cambazoglu, and Hakan Ferhatosmanoglu. 2012. A large-scale sentiment analysis for Yahoo! Answers. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pages 633–642.

Yuming Lin, Jingwei Zhang, Xiaoling Wang, and Aoying Zhou. 2012. An information theoretic approach to sentiment polarity classification. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 35–40.

Bing Liu. 2010. *Sentiment Analysis and Subjectivity*. 2 edition.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Andrew L Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics.

Yelena Mejova and Padmini Srinivasan. 2011. Exploring feature definition and selection for sentiment classifiers. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 546–549.

Prem Melville, Wojciech Gryc, and Richard D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1275–1284, New York, New York, USA. ACM Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*, pages 1–12.

Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2012. Random walk weighting over SentiWordNet for sentiment polarity detection on Twitter. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 3–10. Association for Computational Linguistics.

Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2014. Ranked WordNet graph for sentiment polarity classification in Twitter. *Computer Speech & Language*, 28(1):93–107.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*, pages 70–77, New York, New York, USA.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *ESWC2011 Workshop on Making Sense of Microposts*, pages 93–98.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 19–21.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 271–278.

Bo Pang, Lilian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Empiricial Methods in Natural Language Processing*, pages 1532–1543.

John Pestian, John Pestian, Pawel Matykiewicz, Brett South, Ozlem Uzuner, and John Hurdle. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 5(1):3–16.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

James Spencer and Gulden Uchyigit. 2012. Sentimentor: Sentiment analysis of Twitter data. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 56–66.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Luke Kien Weng Tan, Jin Cheon Na, Yin Leng Theng, and Kuiyu Chang. 2011. Sentence-level sentiment polarity classification using a linguistic approach. In *Proceedings of 13th International Conference on Asia-Pacific Digital Libraries*, pages 77–87.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1031–1040, New York, New York, USA. ACM Press.

Peter Warr, Joanna Barter, and Garry Brownbridge. 1983. On the Independence of Positive and Negative Affect. *Journal of Personality and Social Psychology*, 44(3):644–651.

D Watson, L A Clark, and A Tellegen. 1988. Development and Validation of Brief Measures of Positive and Negative Affect - the Panas Scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070.

Miaomiao Wen, Diyi Yang, and Cp Rosé. 2014. Sentiment analysis in MOOC discussion forums: What does it tell us? In *Proceedings of Educational Data Mining*, pages 1–8.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68.

Theresa A. Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. 2011. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Technical report.