THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

LIWC-UD: Classifying Online Slang Terms into LIWC Categories

Citation for published version:
Bahgat, M, Wilson, SR & Magdy, W 2022, LIWC-UD: Classifying Online Slang Terms into LIWC Categories. in *Proceedings of the 14th ACM Web Science Conference.* Association for Computing Machinery (ACM), pp. 422–432, 14th ACM Web Science Conference 2022
, Barcelona, Spain, 26/06/22. https://doi.org/10.1145/3501247.3531572

Digital Object Identifier (DOI):
10.1145/3501247.3531572

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 14th ACM Web Science Conference

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 24. Sep. 2022

LIWC-UD: Classifying Online Slang Terms into LIWC Categories

Mohamed Bahgat[1], Steve R. Wilson[2], Walid Magdy[1,3]
[1] University of Edinburgh, Edinburgh, UK
[2] Oakland University, Michigan, USA
[3] The Alan Turing Institute, London, UK
m.bahgat@ed.ac.uk,stevenwilson@oakland.edu,wmagdy@inf.ed.ac.uk

ABSTRACT

Linguistic Inquiry and Word Count (LIWC), a popular tool for automated text analysis, relies on an expert-crafted internal dictionary of psychologically relevant words and their corresponding categories. While LIWC's dictionary covers a significant portion of commonly used words, the continuous evolution of language and the usage of slang in settings such as social media requires fixed resources to be frequently updated in order to stay relevant. In this work we present LIWC-UD, an automatically generated extension to LIWC's dictionary which includes terms defined in Urban Dictionary. While original LIWC contains 6,547 unique entries, LIWC-UD consists of 141K unique terms automatically categorized into LIWC categories with high confidence using BERT classifier. LIWC-UD covers many additional terms that are commonly used on social media platforms like Twitter. We release LIWC-UD publicly to the community as a supplement to the original LIWC lexicon.

CCS CONCEPTS

• **Computing methodologies** → **Language resources**; *Neural networks*; • **Information systems** → *World Wide Web.*

KEYWORDS

LIWC, Urban Dictionary, Lexicons, Expansion

ACM Reference Format:
Mohamed Bahgat[1], Steve R. Wilson[2], Walid Magdy[1,3]. 2022. LIWC-UD: Classifying Online Slang Terms into LIWC Categories. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3501247.3531572

1 INTRODUCTION

Linguistic Inquiry and Word Count (LIWC) [29] is a lexicon-based tool. Although LIWC is based on simply providing word counts, it has proven to be widely useful for a range of text analysis tasks. The power of LIWC lies in its ability to measure psychologically relevant dimensions, defined by experts, which capture both linguistic dimensions such as personal pronoun usage and psychological dimensions such as affective and social processes. The dimensions have been validated through a range of studies covering, among others, the measurement of emotions [18], social hierarchies [17], and deception [26]. More recently, LIWC has been used extensively in the domain of web science, including being applied to understand user-generated descriptions of happy moments [15], to measure psychological processes within various types of hate speech comments [12], to generate features for the automatic identifying of political trolls on social media [1], and to explore differences between groups of social media users discussing mental health topics [6]. LIWC has even been translated into several languages including German [23], Chinese [48] and Brazilian Portuguese [7], furthering its usability across languages.

However, reliance on the provided LIWC lexicon (also referred to as LIWC's internal "dictionary") has several key limitations. These limitations include general issues with lexicon-based approaches' handling of phenomena like polysemy and negation. Additionally, while LIWC's lexicon has a specific category for informal language that covers different types of uses such as *Netspeak*, *swear words*, as well as other sub categories that contain words like "lol", "btw", and "thx", we find that many commonly used words on platforms like Twitter are not covered by LIWC's lexicon. This problem is exacerbated by the fact that LIWC is updated, on average, every seven years (the major versions were released in 2001, 2007, 2015 and most recently 2022), which creates problems with coverage of neologisms or evolving word meanings in online discourse. This issue is not limited to lexicon-based methods – even powerful language models degrade in performance over time due to short term changes in language used on social media [16].

To address these and similar problems, lexicon expansion techniques have been proposed. Different work rely on distributional semantics based word embeddings [2, 13] or information retrieval based approaches [47], which require any newly added words to be used a sufficiently large number of times in similar contexts to words already defined in LIWC. This limits their ability to capture rare or very recently popular words.

In this paper, we propose a fundamentally different approach to adding new terms to LIWC lexicon: we use supervised learning to leverage a large, crowd-built online dictionary of online slang terms the Urban Dictionary[1] (UD), to label a large set of terms belonging to many of the core English-language LIWC categories. UD activity has been shown to closely track usage of terms on Twitter, with definitions being created at similar times or even before terms start to trend on the microblogging platform [45]. This fact makes UD an excellent resource for capturing the meanings in a quickly evolving language online. We focus on slang terms as those represent a quickly evolving part of the language and providing resources for these terms are useful for solving relevant NLP tasks [44]. We train

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WebSci '22, June 26–29, 2022, Barcelona, Spain
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9191-7/22/06…$15.00
https://doi.org/10.1145/3501247.3531572

[1] https://www.urbandictionary.com/

a deep learning model to predict the category of a given term based on its definition in UD, using the lexicon within LIWC to generate reference labels for our training and test datasets. We verify our model precision via a static test set, and then apply the model to hundreds of thousands of terms in UD that are currently not covered by LIWC, producing a set of 141,021 new terms that have been assigned a LIWC category with high model confidence. We publicly release this set of terms[2] with the name *LIWC-UD*. The new LIWC-UD lexicon can be added to the core LIWC lexicon in order to gain a substantial increase in coverage for words commonly used on online platforms like Twitter.

2 RELATED WORK

Manual curation of lexicons is a laborious and expensive process. Thus, there has been significant work in trying to create lexicons automatically or add terms to existing ones. These tasks have proven be none-trivial to challenging to obtain accurately expanded lexicons [43]. Badaro et al. [5] created EmoWordNet expanded from DepecheMood [36], a lexicon that was automatically generated from readers annotating news articles with eight emotion categories. Terms in the source lexicon, DepecheMood, were matched in WordNet [24]. The corresponding synonyms (synsets) for these terms were assigned the same emotion labels based on the assumption that synonyms preserve emotions. The resulting new EmoWordNet lexicon was 67K terms compared to 37K for DepecheMood. The original and expanded lexicons were evaluated on the same downstream task with the expanded lexicon outperforming the original one. WordNet was also used by Shaikh et al. [35] to expand the ANEW lexicon [9] which contains 2,477 terms along with their corresponding valence, dominance and arousal. The work merged ANEW with Warriner's lexicon [42] that is derived from WordNet which contained around another 14,000 terms. In one method, synonyms for the combined list of terms were used to get a total of 22,756. In another method, synonyms as well as hyponyms were selected to get a total of 109,752 terms. To validate their method, the authors expanded only ANEW and then used Pearson's coefficient to measure correlation with Warriner's lexicon in addition to validating the resulting terms manually. Khawaja et al. [19] proposed a method for expanding lexicons for specific target domains. The method is based on pointwise mutual information between words in a seed lexicon and selected words from an unlabeled corpus belonging to the target domain. The method was applied to two different lexicons: EmoSenticNet [31]; an automatically generated lexicon with 6 emotion categories with polarity, and NRC Word-Emotion Association Lexicon (EmoLex) [25]; a crowd sourced lexicon with 8 emotion categories in addition to sentiment labels. The resulting lexicons were evaluated on an emotion detection for software developers task [28] where their approach performed consistently better after expanding the two different lexicons.

All of these lexicon expansion methods share limitations. They are either created once then require additional annotations to be updated, or require an unlabeled target corpus that contains a significant number of instances of both the seed lexicon words and any new words to be added to the lexicon. On the other hand, our proposed method relies only on entries added to Urban Dictionary, which has been shown to closely track usage of many newly popular terms and phrases in almost real time on mainstream social media platforms like Twitter [45].

Urban Dictionary (UD) has already been successfully used to build lexicons that cope with ever changing language vocabulary in social media. One example of such work is *SlangSD* [47]. The lexicon contains a total of 96,462 terms and their corresponding sentiment labels. The authors used three methods to annotate words or phrases appearing in UD. First, UD terms were matched to existing lexicons (761 terms). Second, under the assumption that frequent co-occurrence entails similar polarity, new terms which co-occurred with labeled terms in text retrieved from twitter were assigned the same sentiment polarity (another 22,710 terms). Third, for the retrieved terms, synonyms were obtained and labeled with same polarity adding another 72,991 terms. To be able to identify newly added terms, the authors queried UD with specific dates. The resulting lexicon was augmented with the SentiStrength lexicon [38] into SentiStrength_SSD. The new lexicon was evaluated on SMS and Twitter sentiment classification tasks. Both the base and new lexicons were applied via lexicon-based methods for classification Thelwall et al. [39] as well as using a deep learning model. The newly generated lexicon outperformed the two previous ones. In contrast to *SlangSD* which focuses on only sentiment analysis, LIWC-UD is able to capture a wider range of psychologically relevant categories, and relies only on UD only and not any other additional external resources.

Other prior work has approached expanding lexicons as a categorization or term labeling problem. Avancini et al. [4] proposed a method to generate domain sepcific lexicons. Using documents from the target domains, *tfidf* vectors were generated for terms representing terms' contributions to each document. These vectors used as inputs to an *ADABoost.MHKR* [34] classifier to assign target labels. The method was validated by expanding *WordNetDomains* lexicon [21] and then validated on a target domain corpus *RCV1* [20]. While the approach provided a quick way to expand lexicons, the term classification performance was lower when compared to other text categorization tasks. Amir et al. [2] expanded lexicons with more task specific terms. Their work projects generic word embeddings into task specific embeddings using *Non-Linear Subspace Embeddings*; NLSE [3]. These embeddings are generated by training a neural network with a single hidden layer with a cost function that takes into account the labeled terms from the seed lexicon as well as the task targets. The approach was applied to two types of lexicons. One is type is label-based, with varying number of classes such as *Opinion Mining Lexicon* [14], *MPQA* [46], and *EmoLex* [25]. Another type is for lexicons with continuous real numbers corresponding to terms such as *Sem-Lex* [33], *LabMT* [11], *ANEW* [9] and *Ext-ANEW Warriner's Lexicon*. The proposed method was verified by introducing the *NLSE-Lex* sentiment lexicon which was expanded from *Sem-Lex*. *NLSE-Lex* was then used for downstream tasks prediction. The lexicon was applied on three different tasks (one of which had three distinct setups). The results were mostly in favour of the new lexicon, showing the potential of neural network-based approaches to lexicon expansion.

There were also attempts to expand LIWC's lexicon specifically. Empath [13] adds new categories to LIWC using a list of seed words as well as adding new words for existing categories. It starts by

[2]https://github.com/mabahgat/liwc-ud/blob/v2022.1.0.0/LIWC-UD-v1.csv

building non-contextual word embeddings out of a large corpus of fiction stories, tweets, opinions and reviews. The selected corpus is expected to contain more emotional content compared to the usual factual corpora such as Wikipedia. New words are then selected from the embedding space based on their cosine distance from seed words. The newly selected words are further vetted by crowd workers to remove noisy terms. The authors provide 200 new categories that were generated and validated through this processes, which provides a method to obtain new categories on demand. The authors evaluated Empath by measuring the correlation of classification results between Empath and the original LIWC as well as EmoLex and General Inquirer lexicon [37] on 4,500 documents. The average correlation was high ranging between 0.906 and 0.876.

3 RESOURCES

The work presented is based on two resources: LIWC [29, 30] and Urban Dictionary[3].

3.1 LIWC

LIWC[4] is a language analysis tool based on a lexicon that was manually curated by experts in the psychology of language.

LIWC's lexicon comprises of a tree-like structure where there are 11 top-level categories. These parent categories are Function words, Affect, Social, Cognitive Processes (*COGPROC*), Perception (*PERCEPT*), Biological Processes (*BIO*), Drives, Relativity (*RELATIV*), Informal Language (*INFORMAL*), Personal Concerns (*PCONCERN*) and Time Orientation (*timeorient*). For this study, we do not attempt to expand the categories of function words and time orientation (past, present and future tenses) categories as we believe they either can not be expanded or there are better tools for identifying those such as part of speech taggers and parsers. In LIWC's lexicon, a single word can belong to multiple categories, for example, the word *versus* is classified as a member of the function/preposition and cognitive process/differentiation categories.

LIWC employs two matching strategies. One is exact matching for words such as *taxing*. The other strategy employs a pattern based matching to capture more variation of a term such that words with the same prefix are all matched. For example, the LIWC entry *temporar** matches both *temporarily* and *temporary*. In all cases, LIWC is limited to matching only on single words, rather than multi-word expressions.

3.2 Urban Dictionary

Urban Dictionary, UD, is a crowd-sourced resource where users can add terms and their corresponding definitions. Users are also able to up-vote or down-vote definitions that were added by other users. A term can be a single word, multiple words or phrases. Each term definition is composed of a meaning, an example, and a list of tags. These parts of the definition are shown to the web site visitors along with the number of likes and dislikes corresponding to each definition as well as the author and the date when that definition was added.

The content of Urban Dictionary has been shown to be very relevant to language trends. Urban Dictionary was sensitive to discussions on social media [45] where some terms tend to be under focus in Urban Dictionary when these terms trend on social media. A very recent example at the time of authoring this paper is the term *dinobabies* which was added on February 12th, 2022 The term emerged out of a controversy that was first reported in the news at the same day. The definition *dinobabies*[5] defined as "Older workers in the tech industry, formerly known as boomers." was faithful to the actual use of that term in the news.

Table 1: Top 5 terms with highest definition count.

Word	Definition Count
love	1,388
emo	1,382
urban dictionary	1,375
school	1,237
fortnite	1,189

Urban Dictionary's vocabulary is very diverse. While some terms are already included in standard dictionaries, others are newly emerging or being used in fringe communities do not exist in these standard dictionaries. Even for terms that exist in standard dictionaries, Urban Dictionary users can contribute new definitions. For example, the word *man* has meanings that are close to its regular use "not a woman" and "boyfriend, husband, male partner" other definitions refer to its slang use as the following UD definition suggests: "A sentence suffix which, when added, makes anyone sound like the Dude from "The Big Lebowski"". UD also contains definitions for celebrities and political figures such as *Rihanna*, *Obama*, and *Donald Trump* UD also contains definitions for entities such as applications (*instgram*), news outlets (*CNN*), and countries (*USA*). Some Urban Dictionary definitions reflect users' opinions. These opinions represent a more significant percentage of definitions in the case of proper nouns compared to other words [27]. UD terms also include common names. For example, the name *Jennifer* has 173 definitions, most of which are describing specific, yet not well-known individuals. Although these definitions are considered noisy for dictionary purposes and discouraged by Urban Dictionary platform[6] some of these definitions are very popular, with over 1000 more likes than dislikes. Urban Dictionary also includes terms from other languages, such as *haute* (French for *high*), that are usually defined in English.

The data used for our current research was collected for the period of 20 years spanning September 12th, 1999 to December 9th 2019. The total number of definitions collected was 35,349,966 which represent 1,974,244 unique terms with an average of 17.9 definition per term. Table 1 shows the top 5 defined terms with respect to the number of definitions per term. UD defined terms[7] containing more than one word are also frequent in Urban Dictionary. Table 2 shows the frequency of word counts per term. Table 3 shows examples of popular terms of various word lengths.

[3] https://www.urbandictionary.com
[4] Our work used LIWC 2015. More recently, LIWC 22 was published [8].
[5] https://www.urbandictionary.com/define.php?term=dinobabies
[6] The *New Word* page https://my.urbandictionary.com/add.php states "Don't name your friends. We'll reject inside jokes and definitions naming non-celebrities."
[7] Here, *term* refers to either a single word *or* a phrase that is defined in UD.

Table 2: Distribution of number of words per term. Entries in UD define not only single words, but also phrases of varying length.

Gram size	Frequency
unigram	1,048,210
bigram	656,404
trigram	182,366
quadgram	55,437
tetragram+	31,827

Table 3: Most popular entries based on the number of votes for each term length based on number of words.

Term	Definitions	Votes
sex	384	1,484,097
donald trump	585	1,362,214
george w. bush	478	313,589
that's what she said	107	64,182
the cake is a lie	12	32,122
and then i found five dollars	2	31,430
full on double rainbow, all the way	1	24,408

4 CLASSIFYING NEW TERMS INTO LIWC CATEGORIES

To add more entries into LIWC, Urban Dictionary terms are classified into one of the LIWC categories based on their corresponding definitions. To do so, we employed and compared two models; one is the LIWC Max Category Count model and the other uses the transformer-based deep learning model, BERT [10].

4.1 LIWC Max Category Count

Our baseline model is based on counting the number of matches from the original LIWC lexicon. The most frequently matched category in the definition corresponding to a given term t was assigned as the LIWC category label for that term.

All or some of the LIWC categories are specified to be used while counting. Matches on categories other than the specified ones are ignored.

To prevent the model from simply copying the category of t as that term would potentially appear in its own definition (especially in the *example* section of an Urban Dictionary definition), any LIWC pattern that already matches t is discarded while counting category matches within t's definition. That is, if t is *disagree* any matches LIWC entry that match this term such as *disagree** are discarded.

4.2 BERT

BERT [10] is a deep contextualized word embeddings model that generates word representations that are not solely based on the word itself, but also based on words' contexts. In classification tasks, the generated embeddings are usually fed into one (the more common case) or more neural layers. Typically, BERT embeddings are trained on general language data to create base models and then fine-tuned using data from target task. BERT-based models have proven to work well in a multitude of tasks spanning regular NLP tasks such as GLUE [41]; SuperGLUE [40]; and SQuAD [32], and mental and emotional related tasks such as suicide risk assessment [22].

BERT embeddings models come in varying sizes with respect to the number of parameters stored in the model. We choose the uncased version of BERT-base model as it is a compromise between performance in terms of accuracy and required resources to train. BERT-base consists of 12 layers and 110 million parameters. We use the uncased version of the model as the input content is expected to not follow formal casing conventions.

5 TERM CLASSIFICATION SETUP

In order to appropriately handle the hierarchical nature of the classification problem of terms into LIWC's tree-like lexicon categories, the classification of terms into LIWC categories and subcategories is split into multiple subtasks, each with its own model.

The first task is to classify terms based on the root categories. That is, terms are classified between: Affect, Social, Cognitive Processes (*COGPROC*), Perception (*PERCEPT*), Biological Processes (*BIO*), Drives, Relativity (*RELATIV*), Informal Language (*INFORMAL*), Personal Concerns (*PCONCERN*). As mentioned before, the classes of Function words and Time Orientation were not considered for expansion. Next, for each of these categories, a classifier is built to classify terms into the corresponding subcategories. For example, a secondary classifier is trained to classify terms that were labeled previously as belonging to *COGPROC* category by the first classifier into one of *COGPROC*'s subcategories: Insights, Causality, Discrepancy, tentativeness, certainty and Differentiation.

5.1 Data Selection

To obtain labeled data for building our classifier, Urban dictionary terms are matched with LIWC entries. Only unigram terms are considered while matching with LIWC as LIWC 2015 matches only on single word entries. There are two matching strategies that can be used. The first is to use all entries in LIWC, and the second is to restrict matching on exact matches and discard pattern-based "wild card" (e.g., "gambl*") matches. Although wild card matching generates a significant amount of data, the data quality is poor and contains a lot of noise. Thus, we only consider terms that are exact matches to terms in LIWC's lexicon. The selected terms are then split between training and testing sets. The training set is then further split into training and validation sets.

We also filter out stop words and names[8] from our set of data to classify (but not from definitions).

5.1.1 Testing Set. The number of entries that were selected for the test set for the root categories classifier was 1,002. We select the test set first so that we can make sure to exclude LIWC patterns which match test set terms from being used to select any entries for the training set.

For the entire set of terms in Urban Dictionary which are matched by any LIWC lexicon pattern, we compute the percentage that belongs to each of the root categories. These percentages are then

[8]List used to filter names is available at https://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/.

Table 4: Size of test set for each subcategory task.

Category	Size
Affect	249
Bio	125
Cognitive Process	107
Drives	204
Informal Speech	99
Personal Concerns	216
Perception	62
Relativity	140
Social	131

maintained while selecting entries for the test set so that the class distributions remain balanced.

For each of the root categories, the terms out of the 1,002 test set entries that belong to that category are selected. These terms are then further labeled with the appropriate subcategories within that root category. The resulting set is the test set for the classification tasks within each subcategory. Table 4 show the test set sizes for each subcategory task.

The selected Urban Dictionary terms may have more than one corresponding definition. For test set, we only select the top definition based on the difference between likes and dislikes. Also, terms are filtered out if all definitions for the term had more dislikes than likes. Additionally, a term might be tagged by LIWC as belonging to multiple categories. In that case, if the term is labeled by the model as belonging to *any* of these categories, the moodel label will be considered correct.

5.1.2 Training Set. The training set is then created from all the remaining terms that are matched by LIWC but do not match any of the LIWC entries that were used to match terms in the test set. There were four different methods of selecting definitions for the training data that range from focusing on the quality of data to use as much data as possible. Below, we adopt the following notations: the set of terms in the training set is denoted by T_{train}, a term is denoted by t_n, and the set of definitions belonging to a term is denoted by D_{t_n}. We also denote the difference between likes and dislikes as δ such that for a definition d the value is compute as follows:

$$\delta_{d_m} = likes(d_m) - dislikes(d_m) \quad (1)$$

Top-1: Only definitions with the expected highest quality corresponding to each term (similar to the test set selection) are selected. That is, $\forall_{t_n \in T_{train}} |D_t| = 1$. In that case, the number of training examples are much less but the quality of definitions are high.

Top-N: More training definitions are added per term by including the top N if enough definitions exist. That is, $\forall_{t_n \in T_{train}} |D_t| <= N$. Given that we do not define a lower bound for the difference between likes and dislikes δ, some poorer quality definitions are expected to be included in the training set. We select $N = 10$ for our experiments. The value for N was chosen empirically after trying different values.

Table 5: Number of definitions available for training based on each selection method.

Selection Method	Example Count
All	59,907
Top-1	2,279
Top-N $N = 10$	14,462
Min-Diff-P $P = 10$	14,210
Min-Diff-1	25,253

Min-Diff-P: To control the expected quality of the data we restrict the definitions to have a minimum difference between likes and dislikes. That is, $\forall_{t_n \in T_{train}} min(\delta_{t_n}) >= P$. We select $P = 10$. Again, the value for P was selected empirically.

Min-Diff-1: To retrieve the maximum amount of data for our training set we only restrict the included definitions to be voted favorably by *most* of the users regardless of the value of the difference. That is, $\forall_{t_n \in T_{train}} min(\delta_{t_n}) >= 1$.

The number of training samples are shown in Table 5 while Figure 1 shows category class percentages for each case. Note that class percentages vary between the methods. For example, *Top-1* has the highest percentage of terms in *Affect*, *Cognitive Process* and *Relativity* categories while the lowest percentage of terms in *Bio*, *Informal* and *Social*. That is because the number of definitions for each term is different and the quality of definitions with respect to δ varies. So for each selection method, the cut off is made at different percentages of data.

For each training set selection setup a BERT model with a linear layer on top of embeddings output is fine-tuned for 5 epochs while using Adam optimizer with a learning rate of $5e^{-5}$. For validation set, 10% of the training data is used.

5.2 Results

In this section we go through the trained models performances with respect to the different tasks. In all cases the input to the model is presented as a single string and is composed of the concatenation of user-defined tags provided with the definition as space separated words, the definition itself, and usage examples included in the definition if provided.

5.3 LIWC Max Category Count versus BERT

We start by comparing the different model setups on the classification of root LIWC categories; *root-9*. The results on the test set are shown in Table 6. Note that the recall, precision and f-score are computed via weighted average.

Our baseline model, LIWC Max Category count, is clearly performing worse than any of the BERT models. Our top performing model was based on *Top-N*. Doing further analysis on the results of that model, we decided to set a threshold for the model confidence at which a label will be accepted. Figure 2 shows model performance metrics based on precision and recall for our top performing model against different selection thresholds. We favour adding new LIWC entries that are correct rather than adding more entries, thus our focus is to get the highest precision possible. Given that, we selected a confidence threshold of 0.98 to accept annotations for the

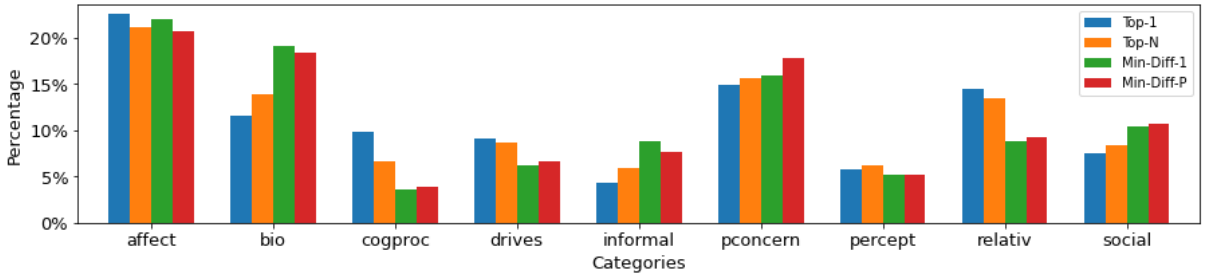

Figure 1: Class distribution for each selection method.

Table 6: Model performance on classifying Root LIWC categories for each data selection method.

Method	Recall	Precision	F-Score
LIWC Max Category Count	0.299	0.336	0.269
Top-1	0.443	0.532	0.452
Top-N	**0.582**	**0.636**	**0.599**
Min-Diff-P	0.561	0.599	0.559
Min-Diff-1	0.417	0.550	0.440

model. Our *Top-N* model was still our best model at that threshold compared to the other models.

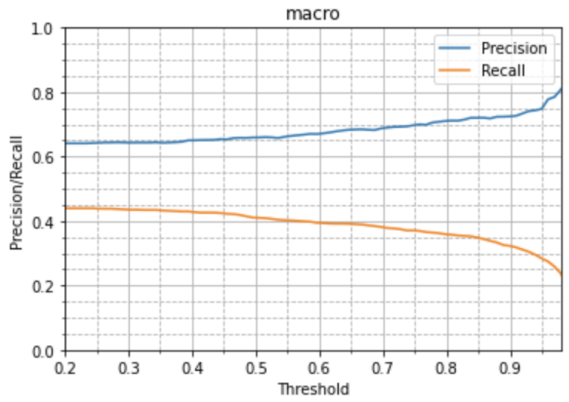

Figure 2: Precision and Recall curves for *Top-N*

5.4 Classifying into subcategories

We used our best data selection method, *Top-N* to select data for each of the subcategories of LIWC that we have included in this work. Figure 3 shows the BERT-based model performance for each of the subcategories. The results show varying accuracy for each subcategory, *Affect* and *Perception* being the subcategories with the best performance. This might be due to the fact that they have fewer subcategories: 2 and 3 respectively, compared to all others. On the other hand, performance on the *Informal* subcategory classification was significantly worse than others. This might be expected given that this category is more ambiguous, referring more to style than content.

6 DISCUSSION

Given the results derived in the previous section, we select *Top-N* for further processing Urban Dictionary terms. In this section we discuss findings from the results.

6.1 Selecting New LIWC Terms

We classify Urban Dictionary terms that were not matched previously in LIWC entries. For each term in LIWC, the definition with the maximum difference between likes and dislikes was selected. To improve the quality of terms being classified and selected into the lexicon, terms were filtered based on different values for the minimum difference between likes and dislikes δ for their corresponding top definition. Figure 4 shows a comparison between three different schemes of selecting a value for δ. Although a value of minimum difference of 10; that is $\delta >= 10$, had by far the lowest amount of terms included, the amount of terms is still significant.

Terms are further filtered based the model confidence. The value selected for the threshold for the terms to be included was 0.98 in order to maintain a high level of precision. The resulting number of terms that were obtained was 141,021. The distribution of categories is shown in Figure 5a.

Next, the terms in the final 141k set are further classified into subcategories. Depending on the category selected for each term, the model corresponding to the category label is used to classify the term between the corresponding subcategories. The same threshold (0.98) for accepting a subcategory label is used. For subcategory classification step, if the confidence is lower than the threshold, the subcategory label is discarded but not the root-level category prediction for the term. The result would be an entry with a root-level category label but without a subcategory label. The resulting term distribution for each subcategory is shown in Figure 5.

The final list of terms is then used to augment the original LIWC lexicon. We refer to the resulting lexicon as *LIWC-UD*.

6.2 Coverage

One way to check the usefulness of the newly added terms is to validate the percentage of text covered by adding the new terms. Regardless of the size of the lexicon, if the content is infrequent there will be little or no use for it. To check our data coverage, we use Twitter as our source of content to check against. We use

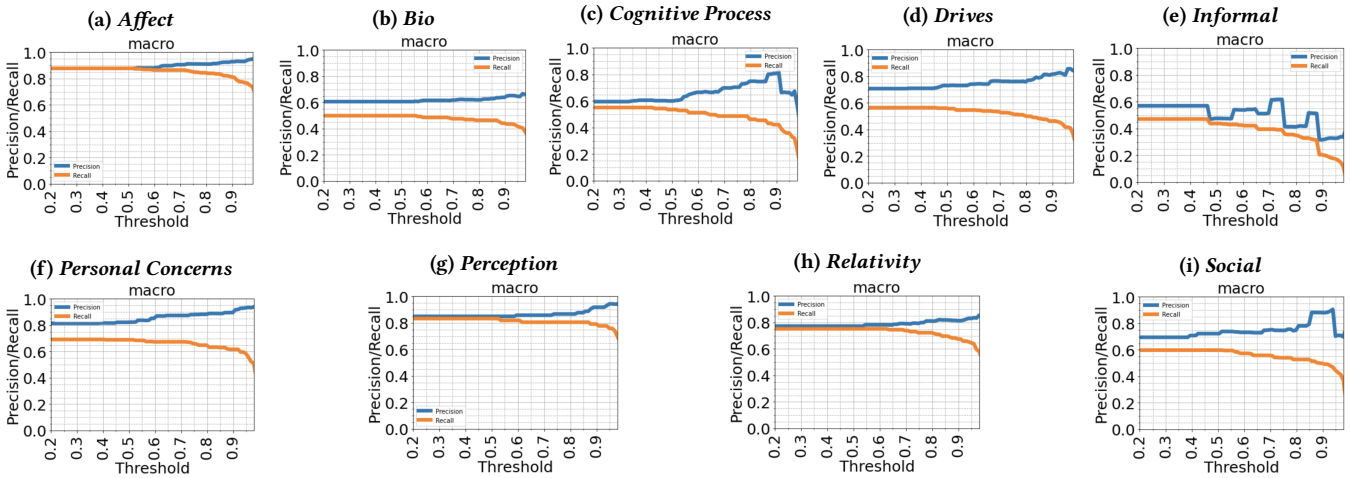

Figure 3: Precision and Recall Curves for LIWC subcategory classification tasks.

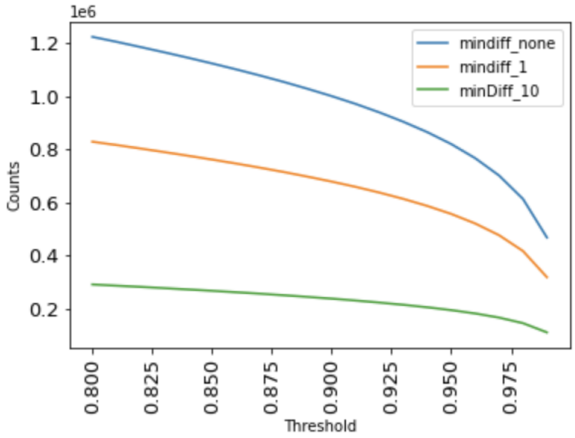

Figure 4: Reduction in the number of newly classified terms for various thresholds and data selection methods.

data from a 1% sample of Twitter between 2011 and 2019. The number of matches was computed for *LIWC-UD* (our resulting lexicon combined with LIWC), original LIWC and LIWC but without wild card matches. Figure 6 compares the percentage of total token matches. *LIWC-UD* has a consistently higher matching percentage in the range of 10% which is almost equivalent to added percentage of LIWC wild cards compared to LIWC exact matches.

Another factor to be looked at is the ability to maintain coverage while the language is evolving. Studying Figure 6, the difference in percentages of words covered widens slightly with time, suggesting that *LIWC-UD* has more coverage of newly adopted terms that were not captured by *LIWC*.

6.3 Multiple Word Terms

One limitation of LIWC 2015 is that it matches single words only. Urban dictionary terms can have multiple words per term. Although our training set contained entries that correspond to single-word terms only (as the source of labeling was LIWC), we used the definitions themselves rather than terms to compute input features for the model. Therefore, the number of words in a term does not impact our ability to classify terms with multiple words into a LIWC category or subcategory. Figure 7 shows the frequency of different n-gram lengths in *LIWC-UD*. Our resulting lexicon has a significant percentage of labeled terms comprised of 2 or more words.

6.4 Analysing Classified New Terms

While machine learning models are well suited for classification problems, lexicons are helpful in analysing content and gaining more insights. In this section, we list example terms from our new lexicon. Table 7 lists samples along with their corresponding output for root category classification and the subcategory classification. We also show for each entry the difference in likes, meaning and example used in the Urban Dictionary definition.

We start by picking randomly from terms that have a relatively high number of votes. These are expected to be more commonly used relative to ones with less votes. Entry #1 and #4; "dracula sneeze" and "crowny" were classified correctly for both the category and the subcategory. For #2 "get an inbox" with $\delta = 8333$ is classified as *Personal Concerns* correctly but the sub category classifier did not output a label with confidence high enough to be included. Entry #3 "textrovert" was classified as *Drives* with subcategory *Affiliation*. The category label was correct. Although it can be argued that the more accurate subcategory would be *Power* according to the first half of the meaning, but *Affiliation* would make more sense based on the second part of the definition as affiliation is expressed through communication of emotions.

Next, we randomly pick terms with relatively low likes difference. For entry #5 "bacongasm", both classifier classified the term

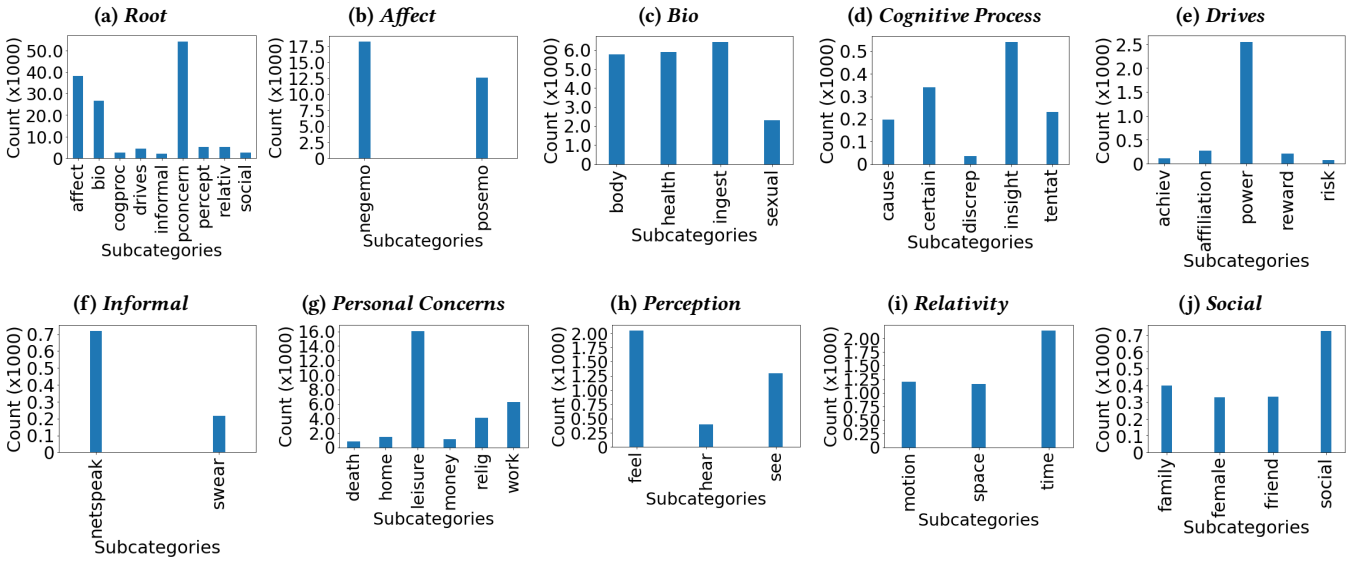

Figure 5: Term count per category

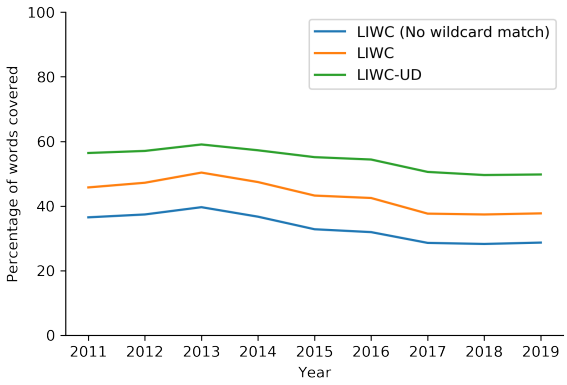

Figure 6: Percentage of total tokens (ignoring stopwords) from a 1% of Twitter covered by each lexicon by year.

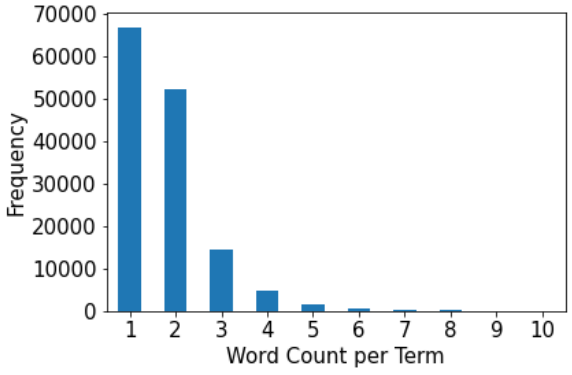

Figure 7: Frequency of N-grams per dictionary entry

correctly for root category and subcategory. The entry #6 "dafting" though was classified incorrectly by the root classifier as *Perception*. That might be due to the presence of words related to listening perception such as "radio" in the example and "music".

Finally, we look at *Informal* class which is the lowest performing sub categories classifier. Entries #7, #8, and #9, "yaab", "hyung" and "deeked" are acceptably classified into *Informal*.

7 CONCLUSION AND FUTURE WORK

In this work we presented a method for adding new terms to LIWC, a popular lexicon used for text analysis. Our method uses machine learning to classify input terms into LIWC categories based on the terms' definitions that were uploaded by internet users on Urban Dictionary. Our approach allowed us to add a significant number of new terms with high precision, and allows us to categorize neologisms used on social platforms with little cost. The proposed method also provides a potential for expanding others lexicons than just LIWC.

Future work should focus on improving recall of the term classifications while maintaining the current high level of precision in order to add even more new terms along with focusing on categories that currently have relatively lower representation in *LIWC-UD*. It would also be fruitful to investigate ways to solve the problem of ambiguity given Urban Dictionary provides us with examples that can allow us to identify the appropriate contexts the words are used in.

REFERENCES

[1] Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13. 15–25.

[2] Silvio Amir, Rámon Astudillo, Wang Ling, Paula C Carvalho, and Mário J Silva. 2016. Expanding subjective lexicons for social media mining with embedding

Table 7: Classification Examples

#	Term	Category	subcategory	Likes δ
1	dracula sneeze	Bio	Health	9,600
	Meaning: holding your arm up over your face in a position similar to Dracula holding up his cape and then sneezing into you elbow			
	Example: Do to the H1N1 swine flu pandemic the Centers For Disease Control recommends using the Dracula sneeze technique to avoid spreading germs.			
2	get an inbox	Personal Concerns	-	8,333
	Meaning: Derived from the expression "get a room." When couples constantly leave romantic, mushy or suggestive messages on each other's Facebook wall for everyone else to see, someone else may tell them to "get an inbox already" and carry on in private.			
	Example: Sadie: I love you SO F****G MUCH ahhhhhhh Im gonna die from how much I love you <3 <3 <3 Ian: me too bby Sabrina: holy s**t guys, get an inbox			
3	textrovert	Drives	Affiliation	6,993
	Meaning: 1. One who feels an increased sense of bravery over texting, as opposed to in person. 2. One who will often only say what they really feel over text messages.			
	Example: Kelly: "So how'd the conversation go with Bill last night?" Wendy: "Ah he's such a textrovert. We didn't make any progress until I went home and he spilled his guts over texts."			
4	crowny	Drives	Power	6,421
	Meaning: Adj. That which is suitable for royalty or someone deserving of a crown. Something highly desirable, of quality or of fine origin. Typically, anything that is the shit, the bomb, killer, dope, etc.			
	Example: I converted my Benz to bio and got a new sound system – now my shit is crowny! That local family farm is so awesome! Their goat cheese is the crowny organies! Question:What's this shit like? Answer: It's crowny, son.			
5	bacongasm	Bio	Ingest	10
	Meaning: The sensation that accompanies a bite into a particularly good piece of bacon. Often occurs with crunching, moaning, and drooling.			
	Example: "Ohh, this is sooo good... I think I just had a bacongasm." Everyone stared as Jeff moaned from his bacongasm.			
6	dafting	percept	-	10
	Meaning: Dancing to Daft Punk music.			
	Example: As soon as that new Daft Punk song came on the radio, everyone stopped what they were doing and began dafting.			
7	yaab	Informal	Netspeak	1,106
	Meaning: pronounced "yawb" Year And A Bit a unit of time measuring one year and a maximum of 11 months.			
	Example: Andrew: "hey, when's your birthday?" Elizabeth: "in a YAAB."			
8	hyung	Informal	787	
	Meaning: Hyung is a word used by korean males to adress another male older than them who they are close to. Hyung literally means "older brother" The female version of this word is oppa.			
	Example: Tomorrow Seungjun hyung and I are meeting up with Jinil hyung.			
9	deeked	Informal	Netspeak	710
	Meaning: An abbreviation for "disqualified." Comes from the acronym DQ, and has been shortened to Deek.			
	Example: How did your race go?" "Not well, I got deeked" "Hey be careful on your start, they might deek you			

subspaces. *arXiv preprint arXiv:1701.00145* (2016).

[3] Ramón Astudillo, Silvio Amir, Wang Ling, Mario J Silva, and Isabel Trancoso. 2015. Learning word representations from scarce and noisy data with embedding subspaces. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1074–1084.

[4] Henri Avancini, Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. 2006. Automatic expansion of domain-specific lexicons by term categorization. *ACM Transactions on Speech and Language Processing (TSLP)* 3, 1 (2006), 1–30.

[5] Gilbert Badaro, Hussein Jundi, Hazem Hajj, and Wassim El-Hajj. 2018. EmoWordNet: Automatic expansion of emotion lexicon using English WordNet. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. 86–93.

[6] Mohamed Bahgat, Steven R Wilson, and Walid Magdy. 2020. Towards Using Word Vector Embeddings Space for Better Cohort Analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[7] Pedro Balage Filho, Thiago Alexandre Salgueiro Pardo, and Sandra Aluísio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

[8] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The Development and Psychometric Properties of LIWC-22. (2022).

[9] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report. Technical report C-1, the center for research in psychophysiology ....

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. 4171–4186.

[11] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one* 6,

12 (2011), e26752.
[12] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[13] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4647–4657.
[14] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 168–177.
[15] Kokil Jaidka, Niyati Chhaya, Saran Mumick, Matthew Killingsworth, Alon Halevy, and Lyle Ungar. 2020. Beyond Positive Emotion: Deconstructing Happy Moments Based on Writing Prompts. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 294–302.
[16] Kokil Jaidka, Niyati Chhaya, and Lyle Ungar. 2018. Diachronic degradation of language models: Insights from social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 195–200. https://doi.org/10.18653/v1/P18-2032
[17] Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology* 33, 2 (2014), 125–143.
[18] Jeffrey H Kahn, Renee M Tobin, Audra E Massey, and Jennifer A Anderson. 2007. Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American journal of psychology* (2007), 263–286.
[19] Hussain S Khawaja, Mirza O Beg, and Saira Qamar. 2018. Domain specific emotion lexicon expansion. In *2018 14th International Conference on Emerging Technologies (ICET)*. IEEE, 1–5.
[20] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research* 5, Apr (2004), 361–397.
[21] Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet.. In *LREC*. 1413–1418.
[22] Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. 39–44.
[23] Tabea Meier, Ryan L Boyd, James W Pennebaker, Matthias R Mehl, Mike Martin, Markus Wolf, and Andrea B Horn. 2019. "LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE-LIWC2015. *PsyArXiv* a (2019).
[24] George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.
[25] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence* 29, 3 (2013), 436–465.
[26] Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin* 29, 5 (2003), 665–675. https://doi.org/10.1177/0146167203029005010 arXiv:https://doi.org/10.1177/0146167203029005010 PMID: 15272998.
[27] Dong Nguyen, Barbara McGillivray, and Taha Yasseri. 2018. Emo, love and god: making sense of Urban Dictionary, a crowd-sourced online dictionary. *Royal Society open science* 5, 5 (2018), 172320.
[28] Marco Ortu, Alessandro Murgia, Giuseppe Destefanis, Parastou Tourani, Roberto Tonelli, Michele Marchesi, and Bram Adams. 2016. The emotional side of software developers in JIRA. In *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*. IEEE, 480–483.
[29] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
[30] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.
[31] Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems* 69 (2014), 108–123.
[32] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.
[33] Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment Analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, 451–463. https://doi.org/10.18653/v1/S15-2078
[34] Fabrizio Sebastiani, Alessandro Sperduti, and Nicola Valdambrini. 2000. An improved boosting algorithm and its application to text categorization. In *Proceedings of the ninth international conference on Information and knowledge management*. 78–85.
[35] Samira Shaikh, Kit Cho, Tomek Strzalkowski, Laurie Feldman, John Lien, Ting Liu, and George Aaron Broadwell. 2016. ANEW+: Automatic expansion and validation of affective norms of words lexicons in multiple languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 1127–1132.
[36] Jacopo Staiano and Marco Guerini. 2014. Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 427–433.
[37] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis. (1966).
[38] Mike Thelwall. 2017. The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength. In *Cyberemotions*. Springer, 119–134.
[39] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology* 61, 12 (2010), 2544–2558.
[40] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537* (2019).
[41] Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
[42] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45, 4 (2013), 1191–1207.
[43] Steven Wilson, Walid Magdy, Barbara McGillivray, and Gareth Tyson. 2021. Embedding Structured Dictionary Entries. Association for Computational Linguistics.
[44] Steven R Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang NLP applications. ACL.
[45] Steven R Wilson, Walid Magdy, Barbara McGillivray, and Gareth Tyson. 2020. Analyzing temporal relationships between trending terms on twitter and urban dictionary activity. In *12th ACM Conference on Web Science*. 155–163.
[46] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*. 347–354.
[47] Liang Wu, Fred Morstatter, and Huan Liu. 2018. SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation* 52, 3 (2018), 839–852.
[48] Xiangkai Zeng, Cheng Yang, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Chinese LIWC lexicon expansion via hierarchical classification of word embeddings with sememe attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

A LIWC-UD VERSUS LIWC-22

At the time of conducting our experiments and authoring this publication, LIWC 2015 [29] was the most recent version for LIWC. More recently a newer version of LIWC was released under than name LIWC-22 [8]. In this appendix, we provide some details the newest version LIWC-22 and analyse if any of the newly added terms to it has been predicted by our LIWC-UD. This is an opportunity to compare our automated process for expanding a lexicon to that of the experts'.

A.1 Comparing LIWC-22 with LIWC 2015

LIWC-22 has significant updates compared to LIWC 2015. Some new categories were added and others were modified or removed. An example of a new parent category is *Culture* with new subcategories *Politics*, *Ethnicity*, and *Technology*. Categories that were removed are *Comparison words*, *interrogatives* and *relativity* along with some punctuation categories. *Personal Concerns* category was renamed to *Life Style* but kept the same semantics and subcategories. Other categories were split. Examples of those are *Positive emotions* subcategory from LIWC 2015 was split into *Positive Tone* and *Positive Emotion* in LIWC-22 while *Negative emotions* subcategory was split into *Negative Tone* and *Negative Emotion*. Also, *Health*

subcategory in LIWC 2015 got split into further sub-subcategories: *illness*, *wellness*, *mental* and *substance* in LIWC-22.

One of the most important changes in LIWC-22 is the introduction of new terms, which aligns to the goal of our research here. The total number of unique terms in the lexicon was almost doubled in LIWC-22 to reach 12,400 terms compared to only 6,547 terms in LIWC 2015. This expansion affected all categories. For example, *Cognitive Process* category has 1,365 terms in LIWC-22 compared to 797 in LIWC 2015. Also, *Visual* has 226 terms in LIWC-22 compared to 126 in *See* category in LIWC 2015.

Another significant change in LIWC-22 that addressed one of LIWC 2015 limitations mentioned in our work is the inclusion of multiple word terms in the lexicon. There are terms that have two or more words. Examples are: "her fault" and "civil unrest" for bigrams, "not so much", "loss of appetite" and "over the moon" for trigrams and "not in the mood" and "not by any means" for quadgrams.

A.2 Assessing LIWC-UD against LIWC-22

LIWC-UD is based on the expansion of LIWC 2015, and since LIWC-22 included an expansion of around 6000 terms compared to LIWC 2015, we checked if any of those added terms has been already predicted by our LIWC-UD and if their predicted categories matches those assigned by experts in the newer version of LIWC. This was an excellent opportunity to assess the performance of our method to expand LIWC on the sample of terms that has been added to LIWC-22 by experts.

We only considered single-word terms in LIWC-UD for the assessment process. Multiple word terms were excluded because matches from those were found to match on the word level rather than full term matches. Out of the 67,149 single-word terms added by our LIWC-UD, we found 6,861 (10.42%) of those were covered by the new LIWC-22 either as an exact match or wildcard match. This shows that LIWC-UD is still a significant contribution, where its coverage is still superior to the newer version LIWC-22.

Next, we compute the percentage of matching term categories between LIWC-UD and LIWC-22. LIWC-22 category labels were mapped back to LIWC 2015 equivalent. Categories in LIWC-22 with no matching LIWC 2015 categories were discarded. Also, terms that are annotated with mapped LIWC-22 categories are filtered such that only categories that we selected to extend in LIWC-UD are included. The number of remaining terms to asssess was 6,012. Assuming the LIWC-22 labels as the ground-truth, the accuracy of the terms identified by our LIWC-UD was 65% (3,907 out of the 6,012 matched the category assigned by experts in LIWC-22). For subcategories, out of the terms with matched root categories there were 3,090 terms labeled with subcategories (the ones were the corresponding subcategory model's confidence was above the selected threshold). The accuracy on the subcategory level was 76%.

These results shows the power of our methodology for expanding a lexicon such as LIWC. While the accuracy of the expanded terms ranges between 65% to 76% on the category and subcategory levels, the method allows of the expansion of those terms at almost no cost and with order of magnitude larger coverage. Thus, we hope LIWC-UD will be a resource that serves as the original LIWC but at a larger scale. In addition, and most importantly, it can be used as a pool for experts to revise and validate its predictions for newer version of LIWC in a step to speed and minimise the cost of the process.