

Lexpansion: Evaluating Dictionary Based Lexicon Expansion for Social Media Analysis

Mohamed Bahgat¹, Steven R Wilson², Walid Magdy¹

¹University of Edinburgh

²University of Michigan-Flint

m.bahgat@ed.ac.uk, steverw@umich.edu, wmagdy@inf.ed.ac.uk

Abstract

Lexicons are indispensable tools for textual analysis through labelled term associations. Despite their utility, lexicons are static and require manual effort to curate and maintain. Regular updates are essential to stay relevant amid semantic shifts and neologisms during language evolution. In this work, we explore the potential of supervised learning for expanding lexicons using dictionaries. We study the effect of using dictionaries with varying properties such as noise, size, labels, structure and curation method. Definitions are used as input features to a transformer model (BERT) that assigns categories to terms. We analyse the expansions using varying English dictionaries and lexicons for estimated accuracy, coverage and labelling consistency and apply the expanded versions to a downstream task. Our analyses show dictionary based expansion is a robust approach. We release our expanded lexicons, code, and pretrained models.

1 Introduction

Lexicons have been broadly used for text analysis in domains such as politics (Elbagir and Yang 2019), social attitudes (Gao and Huang 2017), mental health (Zanwar et al. 2023) and more. Even in the era of LLMs, lexicons are useful to interpret results, build lightweight models or add emphasis on relevant features (Li et al. 2020).

However, lexicons are static, yet language evolves over time. Various societal and demographic factors were shown to affect language change (Labov 2011; Milroy and Milroy 2017). These changes are amplified by social media both syntactically (Maity et al. 2016) and semantically (McGillivray et al. 2022). Only a limited number of lexicons are updated to cope with language changes, like LIWC, (Pennebaker, Francis, and Booth 2001; Pennebaker et al. 2007, 2015; Boyd et al. 2022) which has been revised at 7-year intervals. An alternative is using resources such as dictionaries to expand lexicons (Bahgat, Wilson, and Magdy 2022). Dictionaries are updated frequently and capture language changes (Nguyen, McGillivray, and Yasseri 2018). For example, in March 2023, Oxford English Dictionary (OED) added 700 new terms like “deepfake” and “groomzilla” while revising 1,400¹. Other dictionaries are

even more sensitive. In Urban Dictionary (UD), a crowd-sourced dictionary, “selfie” was first defined in September 2009 versus August 2013 for OED (Nguyen, McGillivray, and Yasseri 2018). UD contains a variety of slang terms and is useful in relevant applications (Wilson et al. 2020a).

Downstream tasks that analyse content from social media would benefit largely from such increased coverage especially for terms that commonly surface on these forums.

In this work, we present “Lexpansion”, a framework for adding new terms to lexicons by obtaining candidates from dictionaries and analysing results. We specifically address categorical lexicons which comprise lists of terms and their corresponding categories. Dictionaries used in our work are lists of words and corresponding senses. We describe senses as definitions composed of meaning and any examples provided for a given term. Dictionary definitions that correspond to lexicon terms are used to train a model to classify other unseen dictionary terms into one of the categories. Models are evaluated on held-out lexicon terms and their corresponding dictionary definitions.

Our work focuses on understanding the effect lexicon and dictionary properties have on dictionary-based lexicon expansion. We propose a framework to (1) train a model once then extract candidates from batches of data without the need to retrain the model, (2) compare results from different dictionaries for lexicon expansion, (3) check category assignments for the same and across different dictionaries for multiple definitions for the same term, and (4) study the impact of using expanded lexicons on downstream tasks. Our method targets lowering lexicon update costs and enables adaptation to language shifts with affordable, accessible models. We discuss English resources, but given our work does not rely on language specific features, it applies to other languages. Our code, models and expanded lexicons are available publicly².

2 Related Work

Lexicons are used in sociolinguistics and computational social science including analysing political speeches (Liu and Lei 2018), elections discussions on social media (Tumasjan et al. 2010), hate speech (ElSherief et al. 2018), social hierarchies and cohorts (Kacewicz et al. 2014), mental health

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://www.oed.com/information/updates/march-2023/>

²<https://github.com/mabahgat/lexpansion>

disorders (Coppersmith, Dredze, and Harman 2014), and emotions versus values (Bahgat, Wilson, and Magdy 2020). Lexicons are used as additional features with machine learning models (Biggiogera et al. 2021; Cutler et al. 2021), augmented during training neural layers for domain specific information (Lei, Yang, and Yang 2018; Li et al. 2020) or emphasize features from a target task (Baziotis et al. 2018).

2.1 Early Automated Lexicon Creation and Expansion

Neologisms often emerge in response to cultural, technological, or social changes (Čilić and Plauc 2021), which poses challenges to lexicons’ relevance, including within non-English languages (Bies et al. 2014; Cassidy et al. 2022). Generating or expanding lexicons automatically provides a solution. For example, for morphological lexicons, Oliver, Castellón, and Márquez (2003) and Oliver and Tadic (2004) expanded two morphologically rich languages, Russian and Croatian. Kaufmann and Pfister (2010) attempted to predict morphosyntactic features for out of vocabulary using a statistical model and applied it to German. Watkinson and Manandhar (1999), Thomforde and Steedman (2011) and Wang, Kwiatkowski, and Zettlemoyer (2014) expanded automatically Combinatory Categorical Grammar lexicons. There were also attempts to expand WordNet in French Sagot and Fišer (2012). Similarly, Seppälä, Barque, and Nasr (2012) used rules to generate a semantic lexicon for French adjectives while Cai and Yates (2013) proposed expanding the lexicon used for semantic parsing through pattern matching.

As for categorical lexicons, a common approach is adding synonyms of current entries from resources like WordNet (Miller 1998) as in the case of Badaro et al. (2018) which expanded *DepecheMood* (Staiano and Guerini 2014) creating *EmoWordNet* containing 1.8 times more terms. Similarly, Shaikh et al. (2016) added hyponyms to get 5 times more terms. WordNet is manually curated with high quality terms and relations but suffers from being stale³ unlike dictionaries that we use in this work as source of candidates that gets updated with new terms.

2.2 Dictionaries

Other work used dictionaries for expansion. Bahgat, Wilson, and Magdy (2022) used Urban Dictionary (UD) to expand LIWC2015 categories. The terms were classified by using their corresponding definitions. However, that work is limited to a single dictionary and does not investigate the impact of lexicon and dictionary properties. We obtained better results using Wiktionary with LIWC2015 and only slightly lower results using OPTED, a much smaller dictionary. Similarly, Wu, Morstatter, and Liu (2018) used UD to build *Slang Sentiment Dictionary* (*SlangSD*) lexicon. The approach leverages sentiment entries from existing lexicons as seed terms and assigns sentiment strength to new candidate terms based on their co-occurrence with seed terms in UD social media posts. However, this method focuses solely on sentiment and is susceptible to errors due to co-occurrence

³Latest update was 2006 as mentioned in <https://wordnet.princeton.edu/download/current-version>

reliance. Wiktionary (WK) was also used for creating and expanding lexicons. Bajčetić and Declerck (2022) added English pronunciation information from WK to WordNet to disambiguate terms with the same spelling but different meanings. Sajous et al. (2010) extended the relations in WK. The work computed graphs derived from Bengali WK including translations, synonyms, and glosses to identify synonym relations via random walks through these graphs. Those relations were then manually reviewed. Di Natale and Garcia (2024) identified new terms by leveraging shared senses derived from aligning terms in language-to-language dictionaries. That is, when a term from a source language is mapped to a term from a destination language, all terms in the source language that correspond to the destination language term are considered as candidates. But language-to-language dictionaries are not expected to have the same coverage and fast evolution as dictionaries defining terms in the same language.

2.3 Embeddings and Tailored Features

More approaches rely on word embeddings where terms are represented by vectors. Vectors for semantically similar terms have smaller differences (Mikolov et al. 2013). That was exploited by Empath (Fast, Chen, and Bernstein 2016) to select candidate terms that are neighbouring the existing terms in LIWC lexicon in the embeddings space for manual review. The space was built using fictional text with more emotional content which is more relevant to LIWC, a psychometric lexicon. Similarly, Wilson, Shen, and Mihalcea (2018) created a lexicon by starting with seed terms for personal values and then adding candidate terms neighbouring the seed terms in the embeddings space. Again, the terms were reviewed manually to increase quality.

In other work, tailored features were used to classify terms. (Avancini et al. 2006) expanded lexicons for new domains by computing *tfidf* feature vectors for terms from domain corpora that were fed into *ADABOOST.MHCR* classifier to label them. The resulting lexicon had higher coverage than a test target domain corpus. In another case, Staiano and Guerini (2014) construct a document-by-emotion and word-by-document matrices to estimate the labels for each word to expand *DepecheMood* to create *DepecheMood++*. Those matrices were estimated from web articles where users labelled each with 8 emotions. However, these methods require additional specialized corpora to compute features for new terms rather than only relying on dictionary definitions.

Compared to using dictionaries where we can build a model once and then classify new terms, adding new terms in the above approaches would require rebuilding the embeddings or retraining the classifiers to maintain accuracy.

2.4 Transformers and Large Language Models

Lexpansion assigns class labels to word sequences similar to, for example, Word Sense Disambiguation (WSD) (Bevilacqua et al. 2021). Huang et al. (2019) concatenated context and gloss and fine tuned BERT (Devlin et al. 2019) to predict if the gloss is correct while Yap, Koh, and Chng (2020) predicted the gloss directly by using glosses from WordNet and manually annotated corpora, but new glosses

or words will not be found in WordNet. Also, manually annotated corpora will rarely be updated relative to dictionaries. In another case, ChatGPT was used to annotate terms automatically (Marcondes et al. 2024). These annotations are then reviewed manually. LLMs do a “fair” job in those experiments but still require manual review.

3 Lexpansion

Lexpansion leverages dictionary term definitions as inputs to predict a lexicon category using a classifier. An experiment expanding a lexicon can train multiple models, each with a different dictionary, allowing us to compare and analyse the results. We formulate lexicon expansion as a multi-class classification problem where a term is assigned a lexicon category based on the term’s corresponding definition.

3.1 An Expansion Run

A lexpansion *run* attempts to expand a lexicon. A run can be viewed as a pipeline which takes a lexicon to expand, one or more dictionaries as source of candidates, exclusion lists used to filter out entries in dictionaries, and a configuration file for model parameters. The configuration file includes lexicon categories to be expanded, dictionaries used, training and testing set sizes, which exclusion lists to use, threshold to filter out dictionary entries if applicable, model parameters and model labelling confidence threshold applied when generating the final expanded lexicons. The pipeline steps are processing dictionary entries, creating the training and testing sets, training the model, running the evaluation on the test set, generating a report and optionally running on the full dictionary to generate an expanded version of the lexicon. If the run involved multiple dictionaries, the evaluation report would provide a comparison. Also optionally, output for multiple models trained on different dictionaries can be used in majority voting when assigning labels to terms.

3.2 Lexicons

We expand three lexicons that vary in size and complexity. Appendix A lists example entries.

Values (Wilson, Shen, and Mihalcea 2018) is a lexicon that captures personal values. The authors created an initial hierarchy with seed terms. More candidate terms are proposed by selecting neighbouring terms to seed terms from a word embeddings space. The resulting list of terms are expected to be noisy, so crowd-sourced human annotators verified the terms and their assigned categories. The categories we selected for expansion are “life”, “parents”, “truth”, “religion”, “social”, “feeling-good”, “children”, “animals”, “learning”, “order” and “accepting-others”. These are top level categories that were represented enough in the largest of our selected dictionaries. We applied such constraint to obtain enough training and testing examples.

LIWC2015 (Pennebaker et al. 2015) is an iteration of Linguistic Inquiry and Word Count (LIWC) which measures various psychometric properties in input text. LIWC is a hierarchical lexicon that links terms to categories that

Lexicon	Values	LIWC2015	LIWC22
Count	1,664	6,547	12,400

Table 1: Number of all terms in each lexicon.

represent different properties: Linguistic and Grammar Dimensions; physiological, cognitive, perceptual, and biological processes; drives; time orientation; relativity; personal concerns; and informal language. Under each, there can be subcategories. We selected a subset of the top level categories. These were: “affect”, “social”, “cogproc” (Cognitive Process), “percept” (Perceptual Process), “bio” (Biological Process), “drives”, “relativ” (Relativity), “pconcern” (Personal Concerns), and “informal”. Function words category with subcategories of personal pronouns, adverbs, numbers and others was left out as new terms can be identified by other means. Note that LIWC2015 uses wildcards indicated by “*” which only appears at the end of a term. For example, “enjoy*” matches “enjoy” and “enjoyable”. Wildcards expand coverage but introduce noise so we opted not to use them. LIWC2015 only matched unigrams.

LIWC22 (Boyd et al. 2022) is the next iteration after LIWC2015. LIWC22 has a revised set of categories. Some categories were added, such as *Culture*, *Personal Concerns* was renamed as *Life Style*, and *Relativity* was deleted altogether. LIWC22 has also increased coverage using nearly twice as many entries compared to LIWC2015 and more flexible wild cards appearing at any place in terms. LIWC22 also includes n-grams as “civil unrest” or “not in the mood”. We expand a subset of LIWC22 top categories: “Culture”, “Conversation”, “Cognition”, “Drives”, “Social”, “Physical”, “Affect”, “Lifestyle” and “Perception”. Similar to LIWC2015 case, we excluded other categories that can be detected by other means. Table 1 compares lexicon term counts.

3.3 Dictionaries

Dictionaries contain a list of terms and their definitions as well as other information. A term belonging to a dictionary can have multiple definitions. A definition can be composed of multiple parts, but in most cases it will have a meaning and example. Parts are concatenated together to form a definition. We used three dictionaries that we detail below.

Urban Dictionary (UD) is a crowd sourced online dictionary. Authors contribute by adding either new *terms* that did not exist or new *definitions* to already existing terms. Definitions are composed of a meaning and usage example(s). Some definitions are faithful such as “The feeling when your heart overflows with the joy of living and being able to just be free and have fun.” for the term “happy.” However, other definitions are noisy; some are opinionated, like “happy : something that no Gen Z’er experiences” or humorous, like “what UD wants you to buy all the time” as a definition of “mug”⁴. UD users are allowed to up-vote or down-vote a definition. Votes help definitions bubble up, but some highly voted definitions are still considered noise. A common case

⁴Referring to mug advertisements on UD website.

Dict.	Terms	Def.	DPT	TPD
UD	1,974,242	3,534,963	1.79	54
WK	1,065,301	1,376,430	1.29	17.4
OPTED	111,616	176,045	1.58	11.2

Table 2: For each dictionary: number of unique terms, number of definitions, Definitions per term (DPT) and Tokens per definition (TPD).

Dictionary	Values	LIWC2015	LIWC22
UD	608	3,725	81,627
WK	964	5,055	194,478
OPTED	670	3,174	15,463

Table 3: Counts of terms matched from each dictionary.

is proper nouns which usually describe a person known to the author with votes based on popularity rather than soundness. UD also has a significant amount of offensive content (Nguyen, McGillivray, and Yasseri 2018) and contains slang or informal definitions and terms (Wu, Morstatter, and Liu 2018) that are commonly used on social media. Therefore, it is sensitive to new terms which surface on social media (Wilson et al. 2020b) or news (Bahgat, Wilson, and Magdy 2022). UD has a significant number of terms that do not exist in formal common dictionaries.

Wiktionary (WK) is another crowd-sourced online dictionary. In contrast to UD, WK is highly moderated so definitions have higher quality and less noise. Definitions are also more detailed, providing some or all of definitions, examples, pronunciation, etymologies, translations and more. For our task, we use meanings and examples only. WK has more content compared to formal dictionaries, but fewer terms and definitions to UD, as shown Table 2.

The Online Plain Text English Dictionary (OPTED)⁵, is a formal English dictionary. Compared to other formal dictionaries it smaller and not updated. Nevertheless, OPTED is a good representation of classical static dictionaries with formal content, less slang, and minimal noise representing the other side of spectrum compared to UD.

Table 2 compares the three dictionaries. Although OPTED is the smallest, OPTED lists terms not found in the other dictionaries as “testamentize” and “mollities” which are defined as “To make a will” and “With a narrow mouth, as the shell of certain gastropods.” respectively. UD has the highest number of terms and definitions, but there are terms that exist in OPTED and WK but not in UD such as “innocuous” which is defined as “Free from crime; pure; innocent.”. Also, Table 3 shows lexicon matches for each dictionary.

Dictionaries are processed to filter out noisy entries. Names and stop words were also excluded. More lists can be used to filter out other terms such as offensive or biased ones although this would depend on the application as in some cases such terms would be needed.

⁵OPTED is based on “The Project Gutenberg text of Webster’s Unabridged Dictionary” which contains the 1913 US Webster’s Unabridged Dictionary. Can be found at <https://www.mso.anu.edu.au/~ralph/OPTED/>

3.4 Selecting Train and Test Samples

Our superset of samples includes all dictionary definitions for terms matched by a lexicon along with the lexicon category for each term. There can be multiple definitions for the same term and each is considered as a separate training sample. That is, dictionary definitions composed of the meaning (and optionally, examples) are used as input to the model while the lexicon categories are the target labels. In case we can have a metric for dictionary definition quality, low quality definitions are excluded before selecting any samples for either the training or testing sets.

The samples are divided into training and testing sets, ensuring that all samples linked to terms assigned to the test set remain exclusively within it, i.e., if a term in the test set has multiple definitions, none of them appear in the training set. The terms for the test set are selected at random and its distribution of labels matches the distribution of labels in the overall available samples. For terms with multiple definitions, if there is a quality metric then the best one is selected otherwise a definition is selected at random.

Next, terms are selected for the training set. The remaining terms and their corresponding definitions are included in the training set. In UD case, only the top 10 definitions in terms of the difference between the up votes and down votes for each term are included in the training set. Definitions with more down votes than up votes are also excluded even if those are in the top 10. This strategy allows us to exclude noisy definitions that are expected in UD.

An experiment for expanding a lexicon can use multiple dictionaries to allow comparing the results from each dictionary. In that case, a single set of terms are selected where all these terms exist in all the used dictionaries. The test set used to evaluate each model built with different dictionary uses the definitions from that dictionary for the selected terms. The training set is curated the same way as a single dictionary.

3.5 Model

We use pretrained BERT (Devlin et al. 2019) “bert-base-uncased” with a linear layer. All models were fine tuned with $5e^{-5}$ learning rate and 5 epochs. Our implementation allows swapping to other model architectures. We chose BERT for its availability, popularity given the relatively good performance, and ability to train on less expensive hardware. However, our overall framework can be used with any supervised text classification model.

3.6 Tuning Precision

The precision of labels assigned by the model can be tuned by studying Precision and Recall against the value of model confidence assigning these labels. Picking a higher confidence threshold will filter out terms assigned labels with confidence scores below that threshold, resulting in fewer terms added to the lexicon, but with improved precision. The threshold can be selected depending on the desired outcome.

3.7 Using Majority Vote

Precision can be further improved by combining label assignments to a term from different dictionary models for

the same lexicon. Two methods were considered: majority agreement and unanimous model agreement, where a new term will not be added unless multiple dictionary models agree on the same class for a given term. This would result in fewer added terms but may improve precision.

4 Experimental Setup

4.1 Baselines

To assess the proposed method, we ran two baselines to answer different questions. One is whether a simple model achieves comparable performance and the other is whether large language models (LLMs) perform better on expansion task without any fine tuning. We chose those baselines as opposed to previous work as previous work either expanded lexicons with a limited categories or does not address evolving language that the dictionaries allow us to solve.

Our simple baseline was **Max Category Count**, following Bahgat, Wilson, and Magdy (2022). Synonyms and relevant words in the definition are expected to have matching labels. A term is assigned a label equal to the most frequent label matched in its definition. To avoid bias, the candidate term is removed from the definition. For example, the word “pie” is defined in WK as “a type of pastry that consists of an outer crust and a filling. the family had steak and kidney pie for dinner and cherry pie for dessert.”⁶. LIWC2015 “pie” labels as “bio”. The definition included the words “kidney”, “dinner” and “desert” which are all “bio” terms while “social” label was matched once for “family” and “relativ” label was matched once for “outer”.

The second baseline uses **ChatGPT** which yielded impressive results on a range of tasks (Bang et al. 2023; Laskar et al. 2023). We sought to explore whether it could serve as a stand-in replacement for our classification approach. ChatGPT was prompted to label terms with LIWC categories given their definition. We experimented with a range of prompts to find the best style which was *ChatGPT 4 with few shot learning, with lexicon name*. Details are in Appendix B.

4.2 Train and Test Sets Splits

LIWC2015 and LIWC22 test set size were 1,000 samples while Values had 200 given its smaller term count. Training set counts are listed in Table 4. From a dictionary standpoint, WK had the most matches with existing lexicon terms, while OPTED had the fewest. From a lexicon standpoint, LIWC22 had significantly more coverage than the others. We opted not to use wildcard matching for LIWC2015 as we saw empirically that it led to a significant number of noisy terms being matched. Values had the fewest matches.

For UD, to build the training set, we used the top 10 definitions. We use up-votes and down-votes added by UD users as a proxy to define quality. While we believe this is a good approximation, it is worth mentioning that votes can reflect popularity like in opinionated definitions especially when dealing with politicians or public figures.

⁶<https://en.wiktionary.org/wiki/pie>

	UD	WK	OPTED
Values	2,198	3,158	1,535
LIWC2015	17,508	25,179	11,765
LIWC22	182,691	263,018	31,691

Table 4: Training samples count for each pair.

Method	Precision	Recall	F-Score
Max Count	0.34	0.30	0.27
ChatGPT	0.43	0.41	0.38
Lexpansion	0.59	0.57	0.57

Table 5: Comparing Lexpansion to baselines for LIWC2015.

5 Results and Discussion

This section discusses our results. All results were computed using “scikit-learn” (Pedregosa et al. 2011) as multi-class classification and are macro-averaged given class imbalance. Results include all outcomes without filtering using model confidence unless mentioned otherwise.

5.1 Results Compared to Baselines

We compare *Lexpansion* against *Max Count* and *ChatGPT with few-shots* to expand LIWC2015 with UD. The results of held out test set are in Table 5. Compared to a simplistic *Max Count* our method performs much better making the cost of fine-tuning worthwhile. Also, while our purpose-built fine-tuned model is much smaller, it performed better than the none-specialized ChatGPT.

5.2 Multi-Dictionary Experiments

Table 6 shows the results for lexicon-dictionary pairs on held out test set. UD models performed better for Values and LIWC22 while lagging behind WK models for LIWC2015. Although the model built using OPTED was trained on much smaller data (Table 3), the model was not far behind compared to models trained using WK, which had the most data. UD was mostly better in the cases of Values and LIWC22, slightly lagging behind WK for LIWC2015.

5.3 Tuning Results and Using Voting

Table 6 shows test set results running our models. Precision increases beyond 0.8 by increasing the confidence

Lexicon	Dict.	Precision	Recall	F-Score
Values	OPTED	0.67	0.67	0.66
	WK	0.68	0.70	0.68
	UD	0.69	0.75	0.71
LIWC2015	OPTED	0.54	0.54	0.53
	WK	0.61	0.59	0.59
	UD	0.59	0.57	0.57
LIWC22	OPTED	0.67	0.65	0.66
	WK	0.76	0.65	0.68
	UD	0.77	0.77	0.77

Table 6: Results for lexicon-dictionary pairs.

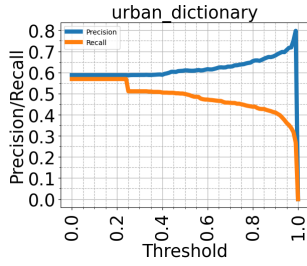


Figure 1: Precision-recall curve for LIWC2015 & UD

Lexicon	Vote	Count	Prec.	Recall	F-Score
Values	2/3	95	0.72	0.69	0.70
	3/3	81	0.89	0.86	0.87
LIWC2015	2/3	246	0.66	0.70	0.67
	3/3	482	0.89	0.95	0.91
LIWC22	2/3	519	0.75	0.84	0.78
	3/3	310	0.95	0.98	0.96

Table 7: Test set results using multiple dictionary models. “Vote” represents agreement in labels while “Count” how many samples are included while computing the metrics.

threshold to 0.98. Figure 1 shows this trade off for expanding LIWC2015 using UD. Figures for the other lexicon-dictionary pairs are in Appendix C.

Another approach, Table 7 shows counts of added terms when two out of three or all three models trained using different dictionaries agree on a label. Note that figures listed in the table were generated without applying a confidence threshold. While precision is higher in these cases, the number of candidate terms is significantly less.

5.4 Expansion Counts

We apply our models to all entries in each dictionary to generate a list of candidate terms for expanding existing lexicons. Table 8 shows the counts for each lexicon-dictionary pair that were assigned labels with confidence more than 0.98. The table also shows the intersection between labels assigned across different dictionaries when only considering the label with the highest confidence.

	Values	LIWC2015	LIWC22
OPTED	9,685	29,543	37,779
WK	163,477	266,021	216,599
UD	354,477	569,009	765,208
Words $UD \cap WK$	15,629	26,008	16,922
Labels $UD = WK$	9,943	13,824	8,622
Words in all	1,723	4,813	4,112
Labels in all	793	2,129	1,398

Table 8: Terms added with model confidence > 0.98 . “Words” show common terms assigned *any label*. “Labels” show common terms assigned *same label*.

5.5 Agreement within Dictionaries

We examine variations in labels assigned to terms having two or more definitions within the same dictionary. Different definitions might still convey similar meanings for a term, which may generate the same label. One example is “semantics” shown in Table 9 where both definitions are related to linguistics. LIWC2015 labels corresponding to each definition were the same “Cognitive Process”. Another example is “wheedle” from OPTED dictionary where both definitions convey similar ideas and were assigned the same LIWC22 label “Social.” “laundry” was also labelled with “life” from Values label with one definition being related to monetary transactions and the other with sanitation as both fall under the parent category “life” through two different subcategories; “wealth” and “health.” In other cases, definitions can convey different meanings, for which models should assign different labels. For example, one definition of “abrade” illustrated a person’s exhaustion, while another described how a physical surface would feel. The corresponding predicted LIWC2015 labels were “affect” and “perception” respectively. Another example is “fam” which relates to family in one definition but to an online store in another. A final example is “gale,” with two definitions; one related to wind and was assigned LIWC22 “Perception” while the other was related to *laughter* and was assigned “Affect”. More examples are in Appendix D. Currently, when generating the final expanded lexicon, only one label is assigned to each term. That label would be the one generated by the model with the highest confidence between the different definitions that correspond to a given term.

5.6 Agreement Across Different Dictionaries

Next, we compare labels assigned by our models for the same term in different dictionaries for the same lexicon. Definitions in different dictionaries can have varying or similar meanings. Table 12 shows definitions for “analytics”. Those definitions conveyed the same meaning thus generating the same LIWC22 “Cognition” label for all dictionaries. In the case of “acai berry”, OPTED did not have an entry while WK had a proper definition, while UD had an opinionated one. The UD definition was still relevant to food and health and was assigned the same label “Physical” (a “food” category in LIWC22) as WK case. For “cheesy”, the models for the three dictionaries assigned the same label “Physical” as those definitions were food related. But there were also definitions in WK and UD that were conveying a sentiment towards a person. In both cases, the term was labelled “Affect” which matches the definitions. A final example is “assailant”. The definitions from each dictionary provide close meanings but with different variations. The model for each dictionary assigned different labels. Note labels assigned by OPTED and WK models match LIWC22 lexicon label, while UD’s does not. Table 10 shows examples of terms that were either matched or different across dictionaries.

We study terms with similar definitions and how consistent their labels are. Considering only terms with two or more definitions, UD had the highest number of definitions per term. However, when calculating the average number of

Term	Lexicon	Dict.	Label	Definition
<i>semantics</i>	LIWC2015	WK	cogproc	a branch of linguistics studying the meaning of words. semantics is a foundation of lexicography.
			cogproc	the individual meanings of words, as opposed to the overall meaning of a passage. the semantics of the terms used are debatable. the semantics of a single preposition is a dissertation in itself.
<i>wheedle</i>	LIWC22	OPTED	Social	To entice by soft words; to cajole; to flatter; to coax.
			Social	To grain, or get away, by flattery.
<i>launder</i>	Values	UD	life	wash scrub clean tidy intermediary To conceal the source of money as by channeling it through an intermediary. Look at us. We're such nerds we're looking up money laundering in the dictionary.
			life	laundrette laundromat clothes An error prone means of removing stains and odor from clothing, often causing them to trade colors, tear, shrink, or to disappear entirely. I accidentally laundered my whites with a red shop rag and now I have pink underwear and socks- I wonder if the missing sock is now pink or white.
<i>abrade</i>	LIWC2015	WK	affect	to wear down or exhaust, as a person; irritate.
			percept	to cause the surface to become more rough.
<i>fam</i>	Values	UD	social	derived from the word "family" . referring to people that are extremely close; as if a family member. what's crackin fam?
			life	Acronym for the Furcadia Alt Market (altmarket.net) Did you see the alts listed on FAM?
<i>gale</i>	LIWC22	WK	Percept	a very strong wind, more than a breeze, less than a storm; number 7 through to 9 winds on the 12-step beaufort scale. it's blowing a gale outside ...
			Affect	an outburst, especially of laughter. a gale of laughter the slightest hint of smugness would have had the nation leaning over our shoulders to blow out the birthday candles with a gale of reproach and disapproval.

Table 9: Candidate terms and their corresponding labels. Note that some definitions are truncated for space.

Lexicon	Term	Top label when using		
		UD	WK	OPTED
Values	enact	life	life	life
	crisis	life	social	children
	nonmated	-	social	-
	clingly	social	social	-
LIWC2015	hiberdating	p. concern	-	-
	leap	relativ	relativ	relativ
	venture	drives	cogproc	relativ
	gaga	affect	affect	-
LIWC22	yonder	Perception	Perception	Perception
	frickle	Drives	Physical	Affect
	mousy	-	Perception	-
	botted	Culture	Culture	-

Table 10: Examples of terms added

unique labels assigned to terms with two or more definitions, the difference between UD and the two dictionaries was small, indicating that definitions in UD often have similar meanings. These values are shown in Table 11.

6 Applying Expansion on Downstream Task

In this section we apply the resulting expanded lexicons on "Cohort Analysis" (Bahgat, Wilson, and Magdy 2020) as a downstream task and analyse the differences in outcomes when compared to running the same using the original lexicons. We aim to study the effect of utilising the expanded lexicon on a non-trivial analysis task.

Cohort Analysis studies how different cohorts affected by mental health challenges relate different concepts. For example, authors might relate "Family" concept to "Home" if

	OPTED	WK	UD
Values	3.42 vs 2.00	3.34 vs 2.10	5.79 vs 2.24
LIWC2015	3.12 vs 2.16	3.17 vs 2.19	6.79 vs 2.51
LIWC22	3.10 vs 2.16	3.36 vs 2.18	8.48 vs 2.47

Table 11: For terms with two or more definitions, average definitions/term versus average unique labels assigned.

they feel family is the most relevant to home while in other cases "Wealth" might be more relevant to "Home" if home is considered as the shelter or source of income. A cohort is represented by the authors of a subreddit. The subreddits included in that study are */r/depression*, */r/SuicideWatch*, */r/mentalhealth*, */r/BPD*, */r/ptsd*, */r/bipolar2*, */r/rapecounseling*, */r/socialanxiety* and */r/StopSelfHarm*. The statistics for each subreddit are listed in Table 13. Consequently, the number of tokens from each subreddit corpus vary between tens of millions to a few millions. Word embeddings modelling the language used by the affected cohort is built using all posts from that subreddit. Those word embeddings were built using fasttext (Joulin et al. 2016) to generate skipgram word embeddings for each subreddit. The vector embeddings size was set to 100. A concept encoded by a lexicon category is represented in the embeddings space by the centroid of all terms included in that category. Relevant concepts are expected to appear close to each other. For a given concept all other concepts are ranked starting with the closest. To avoid ranking concepts that are generally close rather than specifically for a given cohort, ranks are compared to their counterparts in a neutral subreddit */r/IAMA* and reranked according to the difference between how close two concepts

	Dict.	Definition	Label
acai berry/analytically	OPTED	The science of analysis.	Cognition
	WK	the principles governing any of various forms of analysis.	Cognition
	UD	Adjective:A person who is able to analyze and deduce.Noun:To analyze. Adjective:don't worry, he's a analytic.	Cognition
	OPTED	-	-
	WK	a fruit that grows on the brazilian wild palm/bermy tree, euterpe olearacea, used for nutritive support as an antioxidant. it is similar in size to a grape.	Physical
cheesy	UD	berry scam pyramid scheme health fat Acai is a berry native to South America that is pretty healthy, but hasn't been scientifically proven to be any healthier than many other types of berries ...	Physical
	OPTED	Having the nature, qualities, taste, form, consistency, or appearance of cheese.	Physical
	WK	of or relating to cheese. this sandwich is full of cheesy goodness.	Physical
	WK	overdramatic, excessively emotional or clichéd, trite, contrived. a cheesy song; a cheesy movie another night, when the local entertainers had gone home, gould went into the empty lounge to play piano with a cheesy string of colored lights overhead and bongo drums at his side.	Affect
	UD	cheese cheddar mozzarella gouda brie an adjective to describe cheese cheddar is cheesy	Physical
assailant	UD	cheezy sentimental dramatic superficial emotional Sentimental and/or dramatic, yet superficial and unconvincing. The word cheesy is often defined simply as "sentimental", but there is a key distinction between the two. A situation is cheesy when the characters, their motives, and/or their emotions ...	Affect
	OPTED	One who, or that which, assails, attacks, or assaults; an assailer.	Drives
	WK	a hostile critic or opponent. [...] the assailants of the quill have their honour as much at heart as the assailants of the sword.	Social*
	UD	assail ambush waylay attack rob mug A person who attacks someone on a road or a masked attacker. He was attacked by an unknown assailant.	Cognition*

Table 12: LIWC22 models assignments. Labels with "*" got confidence below 0.98 while Dashes "-" got below 0.6.

Subreddit	Submissions	Vocab.	Tokens
depression	766,971	66,661	125,883,951
IAmA	402,415	60,455	22,208,754
SuicideWatch	296,647	44,552	49,142,594
mentalhealth	106,750	31,853	16,177,175
BPD	89,471	26,886	13,686,784
socialanxiety	74,802	22,041	9,674,364
ptsd	21,545	17,042	4,194,729
rapecounseling	14,907	14,488	4,401,250
bipolar2	10,519	9,436	1,424,287
StopSelfHarm	7,965	7,011	1,059,160
Average	179,199	30,043	24,785,305

Table 13: Counts for each subreddit. The data used matches Bahgat, Wilson, and Magdy (2020).

are in a specific subreddit and that neutral subreddit. This is called "relative rank" which is used in comparison moving forward. This task employs two lexicons, LIWC2015 and values, and uses most of the their categories. The content was curated from Reddit, a social media website that allows users to share anonymously. The users use significant slang which is expected to contain new terms not picked up by original lexicons. Thus, we picked up this task to evaluate the downstream performance of the expanded lexicons as (1) it uses more than one lexicon with a wide set of categories and (2) the content curated is expected to contain a lot of slang missing from the original lexicons.

6.1 Coverage

We first compare the coverage of the expanded versus the original lexicons for each subreddit. Coverage is hypoth-

esized to be important as it allows the inclusion of more properties and signals from text leading to a more accurate analysis. For example, if terms relevant to a concept are not matched by the lexicon, it would be under-represented or completely missed in the analysis. The expanded versions of the lexicons include terms added from the three dictionaries mentioned earlier. Table 14 shows the percentage increase for both vocabulary and tokens. For the smaller Values lexicon, the coverage increased significantly for both vocabulary and tokens. While the original LIWC2015 contained more common vocabulary which is apparent given the lower coverage increase in tokens compared to vocabulary, there are missing slang terms which were more commonly used after the lexicon was issued or less frequently used terms given the high cost of including those such as "kabbalah" or the slang "g-d" as in the following excerpt from /r/SuicideWatch "i've been studying *kabbalah* for ages, i'm 18, all i want is to meet *g-d*". Table 15 shows examples of newly matched terms where some can be considered slang such as "meds", "mekill", and "crack".

6.2 Impact

While coverage is higher, but the effect of newly matched terms has to be verified. As such, we investigate the changes in findings when using the expanded lexicons. We empirically chose to filter out labels that were assigned with a confidence less than 0.8. We study how many concepts moved closer or further from others and if those are warranted given the content in the corresponding subreddits. We classify changes into *None*, *minor*, *moderate* and *significant* correspond to zero change, 1 to 4 inclusive, 5 to 9 inclusive and 10 or above respectively. *Minor* changes can be ignored as they can occur because of slight changes while *moderate*

Subreddit	Values %		LIWC2015 %	
	Voc.	Tokens	Voc.	Tokens
bipolar2	+46.52	+86.57	+36.21	+8.72
bipolar	+35.39	+86.77	+31.97	+7.96
BPD	+36.62	+87.6	+32.10	+7.18
CPTSD	+35.25	+86.29	+31.99	+6.99
depression	+23.64	+87.56	+23.36	+5.81
mentalhealth	+30.32	+86.84	+28.16	+6.87
ptsd	+43.98	+86.77	+36.78	+7.13
socialanxiety	+40.76	+87.49	+33.92	+6.07
StopSelfHarm	+55.78	+87.99	+35.91	+5.99
SuicideWatch	+27.75	+87.68	+26.38	+6.11
IAMA	+31.06	+84.82	+29.55	+10.22
Average	+37.01	+86.94	+31.48	+7.19

Table 14: Coverage increase in percentages for each cohort. Tokens are all words appearing in text, while vocabulary is the unique set of words.

Term	Count	LIWC2015 Label
meds	17,674	Biological Process (bio)
online	16,053	Personal Concerns (pconcern)
overdose	7,200	Biological Process (bio)
wasted	6,486	Affect
mekill	4,424	Personal Concerns (pconcern)
noose	3,503	Personal Concerns (pconcern)
disability	3,253	Personal Concerns (pconcern)
crack	1,064	Affect
meth	917	Biological Process (bio)
goodbyes	869	Social
consent	724	Cognitive Process (cogproc)
worthlessness	717	Affect
oblivion	717	Affect

Table 15: Examples of terms matched by the expanded version of LIWC2015 but not the original version along with their counts in “SuicideWatch” and labels.

might be less impactful to the analysis.

Table 16 shows the distribution of changes. An example of newly highlighted findings is the increased relevance of the “Children” concept relative to “Death” among Suicide-Watch authors. This shift suggests a stronger connection between the two topics. Inspection of relevant posts reveals themes of troubled childhoods, exposure to or committing abuse or grief over child loss are linked to early suicidal thoughts. Also, “Affect” which indicates emotions gained relevance compared to “Cognitive Process” which lost relevance, emphasizing that the authors process the concept of “death” more emotionally rather than rationally. Other changes are obvious where categories like “life”, “feeling-good”, “accepting-others” and “social” for both LIWC-2015 and Values had a significant drop in their relevance to “death” for the same subreddit. Those changes show that the extra coverage indeed rendered relevant findings that better aid the analysis task.

Subreddit	Significant	Moderate	Minor	None
bipolar2	1,250	1,084	3,723	915
bipolar	1,269	864	3,746	1,093
BPD	1,294	815	3,813	1,050
CPTSD	1,214	821	3,918	1,019
depression	1,216	784	3,779	1,193
mentalhealth	1,182	818	3,840	1,132
ptsd	1,250	828	3,861	1,033
socialanxiety	1,283	863	3,780	1,046
StopSelfHarm	1,334	976	3,874	788
SuicideWatch	1,179	758	3,792	1,1161

Table 16: Changes in Relative Rank for different subreddits.

7 Conclusion and Future Work

In our work, we presented “Lexpansion” to expand lexicons and analyse with new terms from regularly-updated dictionaries. Dictionary term definitions that exist in the original lexicon are used to train models that label new terms with target lexicon categories. Using Lexpansion, we expanded three lexicons using three dictionaries which vary in properties like formality, sensitivity to new terms, term count, and number of definitions per term. Our analysis of the expanded lexicons shows the effectiveness of Lexpansion. Dictionary-based expansion was also demonstrated to be more effective than other baselines including simple voting and ChatGPT-based while applying the expanded lexicons to a downstream task rendered new findings.

For future work, we aim to identify metrics to quantify parameters like sensitivity to language shifts, ambiguity, and coverage as well as applying our method to multiple languages. We also plan to expand other types of lexicons such as WordNet using similarity of dictionary definitions. Moreover, while currently a single label is assigned to each term, we plan to study assigning multiple labels for terms that their definitions represent multiple senses.

References

- Avancini, H.; Lavelli, A.; Sebastiani, F.; and Zanolini, R. 2006. Automatic expansion of domain-specific lexicons by term categorization. *ACM Transactions on Speech and Language Processing (TSLP)*, 3(1): 1–30.
- Badaro, G.; Jundi, H.; Hajj, H.; and El-Hajj, W. 2018. EmoWordNet: Automatic expansion of emotion lexicon using English WordNet. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 86–93.
- Bahgat, M.; Wilson, S.; and Magdy, W. 2022. LIWC-UD: Classifying Online Slang Terms into LIWC Categories. In *14th ACM Web Science Conference 2022*, 422–432.
- Bahgat, M.; Wilson, S. R.; and Magdy, W. 2020. Towards Using Word Vector Embeddings Space for Better Cohort Analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*.
- Bajčetić, L.; and Declerck, T. 2022. Using Wiktionary to create specialized lexical resources and datasets. In *Proceedings of the thirteenth language resources and evaluation conference*, 3457–3460.

- Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- Baziotis, C.; Nikolaos, A.; Chronopoulou, A.; Kolovou, A.; Paraskevopoulos, G.; Ellinas, N.; Narayanan, S.; and Potamianos, A. 2018. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, 245–255. New Orleans, Louisiana: Association for Computational Linguistics.
- Bevilacqua, M.; Pasini, T.; Raganato, A.; and Navigli, R. 2021. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*, 4330–4338. International Joint Conference on Artificial Intelligence, Inc.
- Bies, A.; Song, Z.; Maamouri, M.; Grimes, S.; Lee, H.; Wright, J.; Strassel, S.; Habash, N.; Eskander, R.; and Rambow, O. 2014. Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 workshop on Arabic natural language processing (ANLP)*, 93–103.
- Biggiogera, J.; Boateng, G.; Hilpert, P.; Vowels, M.; Bodenmann, G.; Neysari, M.; Nussbeck, F.; and Kowatsch, T. 2021. BERT meets LIWC: Exploring State-of-the-Art Language Models for Predicting Communication Behavior in Couples’ Conflict Interactions. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 385–389.
- Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The Development and Psychometric Properties of LIWC-22.
- Cai, Q.; and Yates, A. 2013. Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 423–433.
- Cassidy, L.; Lynn, T.; Barry, J.; and Foster, J. 2022. TwitIrish: a universal dependencies treebank of Tweets in modern Irish. In *60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics (ACL).
- Čilić, I. Š.; and Plauc, J. I. 2021. Today’s usage of neologisms in social media communication. *Društvene i humanističke studije*, 6(1 (14)): 115–140.
- Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 51–60.
- Cutler, A. D.; Carden, S. W.; Dorough, H. L.; and Holtzman, N. S. 2021. Inferring Grandiose Narcissism From Text: LIWC Versus Machine Learning. *Journal of Language and Social Psychology*, 40(2): 260–276.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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.
- Di Natale, A.; and Garcia, D. 2024. LEXpander: applying colexification networks to automated lexicon expansion. *Behavior Research Methods*, 56(2): 952–967.
- Elbagir, S.; and Yang, J. 2019. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 122, 16.
- ElSherief, M.; Kulkarni, V.; Nguyen, D.; Wang, W. Y.; and Belding, E. 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*, volume 12.
- Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657. ACM.
- Gao, L.; and Huang, R. 2017. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*.
- Huang, L.; Sun, C.; Qiu, X.; and Huang, X.-J. 2019. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3509–3514.
- Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Kacewicz, E.; Pennebaker, J. W.; Davis, M.; Jeon, M.; and Graesser, A. C. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2): 125–143.
- Kaufmann, T.; and Pfister, B. 2010. Semi-automatic extension of morphological lexica. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, 403–409. IEEE.
- Labov, W. 2011. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 3. John Wiley & Sons.
- Laskar, M. T. R.; Bari, M. S.; Rahman, M.; Bhuiyan, M. A. H.; Joty, S.; and Huang, J. X. 2023. A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. *arXiv preprint arXiv:2305.18486*.
- Lei, Z.; Yang, Y.; and Yang, M. 2018. Sentiment lexicon enhanced attention-based LSTM for sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Li, W.; Zhu, L.; Shi, Y.; Guo, K.; and Cambria, E. 2020. User reviews: Sentiment analysis using lexicon integrated two-channel CNN-LSTM family models. *Applied Soft Computing*, 94: 106435.

- Liu, D.; and Lei, L. 2018. The appeal to political sentiment: An analysis of Donald Trump’s and Hillary Clinton’s speech themes and discourse strategies in the 2016 US presidential election. *Discourse, Context & Media*, 25: 143–152.
- Maity, S. K.; Ghuku, B.; Upmanyu, A.; and Mukherjee, A. 2016. Out of vocabulary words decrease, running texts prevail and hashtags coalesce: Twitter as an evolving sociolinguistic system. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 1681–1690. IEEE.
- Marcondes, F. S.; Gala, A. d. C.; Rodrigues, M.; Almeida, J. J.; and Novais, P. 2024. Lexicon Annotation with LLM: A Proof of Concept with ChatGPT. In *International Conference on Hybrid Artificial Intelligence Systems*, 190–200. Springer.
- McGillivray, B.; Alahapperuma, M.; Cook, J.; Di Bonaventura, C.; Penuela, A. M.; Tyson, G.; and Wilson, S. 2022. Leveraging time-dependent lexical features for offensive language detection. In *Proceedings of the 1st Workshop of Ever Evolving NLP, EMNLP 2022*.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
- Miller, G. 1998. *WordNet: An electronic lexical database*. MIT press.
- Milroy, J.; and Milroy, L. 2017. *Varieties and Variation*, chapter 3, 45–64. John Wiley & Sons, Ltd. ISBN 9781405166256.
- Mishra, S.; Khashabi, D.; Baral, C.; Choi, Y.; and Hajishirzi, H. 2022. Reframing Instructional Prompts to GPTk’s Language. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 589–612. Dublin, Ireland: Association for Computational Linguistics.
- Nguyen, D.; McGillivray, B.; and Yasseri, T. 2018. Emo, love and god: making sense of Urban Dictionary, a crowd-sourced online dictionary. *Royal Society open science*, 5(5): 172320.
- Oliver, A.; Castellón, I.; and Márquez, L. 2003. Use of internet for augmenting coverage in a lexical acquisition system from raw corpora: application to russian. In *Proceedings of the Workshop IESL 2003. International Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages*.
- Oliver, A.; and Tadic, M. 2004. Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora. In *LREC*.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.
- Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of LIWC2015. Technical report.
- Pennebaker, J. W.; Chung, C. K.; Ireland, M.; Gonzales, A.; and R., B. 2007. The development and psychometric properties of LIWC2007. 1–22.
- Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001): 2001.
- Sagot, B.; and Fišer, D. 2012. Automatic extension of WOLF. In *GWC2012-6th International Global Wordnet Conference*.
- Sajous, F.; Navarro, E.; Gaume, B.; Prévot, L.; and Chudy, Y. 2010. Semi-automatic endogenous enrichment of collaboratively constructed lexical resources: Piggybacking onto wiktionary. In *International Conference on Natural Language Processing*, 332–344. Springer.
- Seppälä, S.; Barque, L.; and Nasr, A. 2012. Extracting a semantic lexicon of french adjectives from a large lexicographic dictionary. In *Joint Conference on Lexical and Computational Semantics*.
- Shaikh, S.; Cho, K.; Strzalkowski, T.; Feldman, L.; Lien, J.; Liu, T.; and Broadwell, G. A. 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.
- Staiano, J.; and Guerini, M. 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.
- Thomforde, E.; and Steedman, M. 2011. Semi-supervised CCG lexicon extension. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1246–1256.
- Tumasjan, A.; Sprenger, T.; Sandner, P.; and Welpe, I. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media*, volume 4, 178–185.
- Wang, A.; Kwiatkowski, T.; and Zettlemoyer, L. 2014. Morpho-syntactic lexical generalization for CCG semantic parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1284–1295.
- Watkinson, S.; and Manandhar, S. 1999. Unsupervised lexical learning with categorial grammars. In *Unsupervised Learning in Natural Language Processing*.
- White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- Wilson, S. R.; Magdy, W.; McGillivray, B.; Garimella, K.; and Tyson, G. 2020a. Urban dictionary embeddings for slang NLP applications. *ACL*.
- Wilson, S. R.; Magdy, W.; McGillivray, B.; and Tyson, G. 2020b. Analyzing temporal relationships between trending terms on twitter and urban dictionary activity. In *12th ACM Conference on Web Science*, 155–163.

Wilson, S. R.; Shen, Y.; and Mihalcea, R. 2018. Building and Validating Hierarchical Lexicons with a Case Study on Personal Values. In *International Conference on Social Informatics*, 455–470. Springer.

Wu, L.; Morstatter, F.; and Liu, H. 2018. SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*, 52(3): 839–852.

Yap, B. P.; Koh, A.; and Chng, E. S. 2020. Adapting BERT for Word Sense Disambiguation with Gloss Selection Objective and Example Sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 41–46.

Zanwar, S.; Wiechmann, D.; Qiao, Y.; and Kerz, E. 2023. SMHD-GER: A Large-Scale Benchmark Dataset for Automatic Mental Health Detection from Social Media in German. In *Findings of the Association for Computational Linguistics: EACL 2023*, 1496–1496.

A Lexicons Category Examples

Tables 17, 18 and 19 shows example terms from selected categories. Note that in the case of LIWC2015 and LIWC22 lexicons, there are wild cards used and denoted using “*” that would allow matching any characters inserted at the corresponding position. In LIWC2015 cases, those wild cards can only appear at the end, while in LIWC22 wild cards can appear in the middle as well.

B Labelling Candidate Terms with ChatGPT

A challenge though when working with LLMs is to find the prompt that would result in the best outcome possible. Asking for the same information using different questions can result in different answers some of which are less accurate or do not provide the required information. For example, adding more content to the question text can cause ambiguity. When enquiring about the value of GDP in the US through ChatGPT through the question “How is the economy doing in the US in terms of GDP value?” would return the answer “I don’t have real-time data, so I can’t provide the current GDP value for the U.S. It’s a dynamic figure that changes regularly. To get the latest information, you might want to check reliable economic sources or government reports for the most accurate and up-to-date GDP data.” which doesn’t really provide the value. But when a shorter more to the point question is used “What is the GDP of US?”, the answer provided “I don’t have real-time data, but as of my last update in 2022, the GDP of the United States was around \$21 trillion. For the most accurate and up-to-date information, you might want to check the latest reports or official sources.” gives the value requested.

To overcome similar issues with querying LLMs, there is work targeting how best to query LLMs for outcomes that correctly answers the questions in hand. One of that work described *prompt patterns* that are analogous to software patterns (White et al. 2023). The work proposed 15 different patterns that addressed a range of cases from question refinement, generating list of facts or infinite generation. Each pattern would have a name, intent and motivation and

Category	Examples		
Life	liv	reside	lifetimes
Parents	daddy	mothers	grandparents
Truth	loyal	faithfully	reality
Religion	prayers	obedience	worships
Social	charity	respectful	relatives
Feeling-good	fervor	joyful	relish
Children	teenager	sons	teenage
Animals	zebras	zoos	iguanas
Learning	lectured	school	university
Order	easily	reassurance	simplicity
Accepting-others	permissive	condone	patient

Table 17: Examples from selected categories in Values.

Category	Examples		
Affect	defense	grr*	badly
Social	grandpa*	cc	flirtatious
Cognitive Process	guessing	wonder	perceiv*
Perceptual Process	cooling	yell	grips
Biological Process	spit	popcorn	wellbeing
Drives	dependent	tentativ*	lacking
Relativity	stuck	ate	emailed
Personal Concerns	ambitions	mecca	apartment*
Informal	4ev*	whoo	haha*

Table 18: Examples from selected categories in LIWC2015 lexicon. The asterisk (*) symbol denotes the presence of a wildcard for that term.

Category	Examples		
Culture	email*	european	veto
Conversation	btw	ttyl	yeah
Cognition	know	but	afaik
Drives	own	approval	regime
Physical	gym*	fish	arm
Lifestyle	game*	hiking	work
Perception	look* for	dizzled	out

Table 19: Examples from selected categories in LIWC22. The “*” symbol denotes a wildcard.

would provide a structure, examples and how the pattern’s use affects and improves the output. Another approach is to *reframe* (or rephrase) a given prompt in a way that would give more desirable results (Mishra et al. 2022). Reframing a prompt would include five key items: (a) avoiding abstract concepts and replacing those low-level patterns that describe the desired output, (b) avoiding descriptive paragraphs and use itemised lists instead while avoiding negations and replacing them with affirmative statements, (c) transforming a larger task into smaller and simpler set of tasks, (d) adding details that would constraint the output, and (e) being specific about the instructions without adding extra details.

While these items mainly target conversational scenarios, we have applied some within our experiments even though our queries were in question-and-answer format⁷. The con-

⁷Using the *completion* API for ChatGPT.

tent of the prompts were stripped to the minimum number of words as long as it remains correctly constructed fluent sentences. Also, the expansion task in itself was reframed into a word labeling task with a single term being presented to the model per query.

There are still different choices to be made on how to format a prompt. For one, there was a choice of whether the low-level detail of which lexicon the target labels stated in the prompt belong to. We hypothesised that if the model has information about the lexicon it would benefit the accuracy of the answer. So, ChatGPT 3.5 was queried about the three lexicons in our experiments. The answers showed that it the model only had information about LIWC2015 but not LIWC22 nor Values. Note that ChatGPT 3.5 training data contains content up to September 2021. Only the most recent ChatGPT 4 model; *gpt-4-1106-preview* had training data with content up to April 2023 (*gpt-4* training data had only up to September 2021). It was able to identify LIWC22 correctly, but still not Values lexicon. Thus, and to optimize cost, when experimenting for baseline performance and whether to include the lexicon name or not, only LIWC2015 mention was added to the prompts and we only tested that baseline on expanding LIWC2015.

Another choice was how to provide the definition within the prompt. The meaning and example could appear separate in the prompt with each labeled with what it was. The other was to provide the whole definition as a context for the word. A final choice that was to consider to test asking the model to generate only the label within the answer. This adds emphasis on the label and simplifies parsing of the answer as the model sometimes generate more details in the answer that can be irrelevant for our task. Examples for both prompt styles separating meaning and example and adding the whole definition as context are shown in Table 20.

In addition to how the prompt is generated, we also experimented with zero versus few shot learning. To generate prompts for few-shot learning, examples were prepended to the best zero-shot prompt based on the above choices. We tested different number of example terms per selected label. Each label was paired with a single term as opposed to stating different terms in bulk that belong to a corresponding label. This makes the prompt clearer to the model and avoid any word sequence pattern that can be confusing. Examples for the few-shot cases are show in Table 20.

To query ChatGPT API, we used *text-davinci-003*⁸ and *gpt-4-1106-preview*⁹. The first is the most capable model for ChatGPT 3.5 and is designed as a completion model. That is, it returns an answer for the provided query without assuming that this is a part of chat conversation. The second model is the most recent one at the time of writing. We have used *gpt-4-1106-preview* is a snapshot of *gpt-4* where *gpt-4* goes through continuous model upgrades. The snapshot should be available for an extended duration which can be used for reproducing the results. Although it is designed for conversation, but according to API documentation, the model can

also be used for completion requests. Given cost considerations, we started off by testing different prompt styles on a sample of the data against ChatGPT 3.5. The styles that showed promise were then tested against the full test set. Furthermore, the highest scoring prompt style was then used to test against ChatGPT 4.

Table 21 shows results from experimenting the different choices of prompts. The zero shot learning styles used was “Definitions as Context” from Table 20, while few shot learning used was “Few Shot Learning with Definition as Context”. When the lexicon name was not mentioned in the prompt, the name was omitted from the sentence preceding listing the different labels. The prompt style of *ChatGPT 4 with few shot learning, with lexicon name* was then used in further comparisons.

ChatGPT API is a paid service with *davinci* model being relatively expensive. To save cost we have run the baseline against the task of expanding LIWC2015 using UD. We assume that this task will be representative of the other two if the difference between our method and the baselines is significant. Also, when we queried about LIWC2015, ChatGPT responded with information about it. But when queried about LIWC22, could not define it as its data is only up to September 2021. ChatGPT did not have information as well about Values lexicon.

C Precision/Recall Curves

Figure 2 shows the precision and recall curves against model confidence for labels assigned by models trained using entries from OPTED, WK and UD against Values, LIWC2015 and LIWC22 lexicons. To obtain higher precision for newly added terms, new candidate terms can be filtered out based on selecting a threshold to reject labels assigned with model confidence less than that threshold. A side effect to the higher threshold is lower recall; that is, the less number of new candidate terms will be available for use. It is left for the users of Lexpansion method to decide a threshold based on the downstream tasks they aim to use the expanded lexicons for by either picking a lower threshold that would increase the number of candidate terms but add more noise or increasing the threshold to get more quality terms but with significantly less number of new terms.

D Labelling Consistency with Multiple Definitions

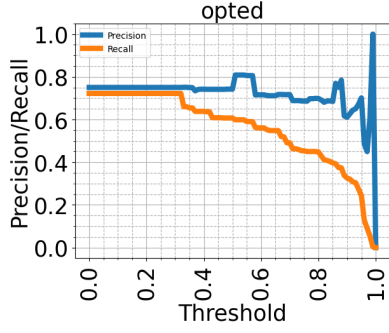
Definitions in UD are commonly repeated or even redundant conveying similar meanings. For example, the term “emo” had 315 definitions labelled by our model with high confidence using UD model while for WK model the same term had only 3. Out of those 315 definitions, 136 contained the word “music”. Note that without filtering out lower confidence labels, the counts would be 1, 382 and 10 for UD and WK respectively out of which 684 had the word “music” in UD’s case. Another example is the word “duh” with 47 definitions in UD and only 2 in WK. One final example is the term “carcolepsy” found only in UD. That term had 5 definitions or a total of 13 when including ones labelled with less

⁸<https://platform.openai.com/docs/models/gpt-3-5>

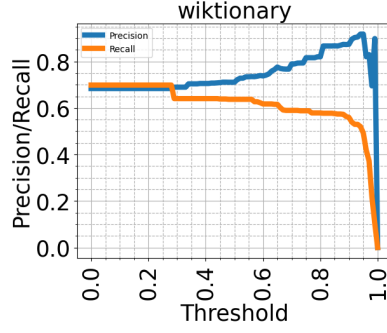
⁹<https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>

Style: <i>Split Definition</i> For a word with the following meaning: “a word commonly used to describe an emotional state in which the person feels a sense of having no hope; usually during a deep depression.” in this example “As I lay awake, alone in my bed, I cannot help but become overwhelmed by this feeling of blah.” which label should be assigned to the word blah from the following “LIWC2015” labels: <i>affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal?</i> Respond with label only
Style: <i>Definition as Context</i> In the following context: “a word commonly used to describe an emotional state in which the person feels a sense of having no hope; usually during a deep depression. As I lay awake, alone in my bed, I cannot help but become overwhelmed by this feeling of blah.” which label should be assigned to the word blah from the following “LIWC2015” labels: <i>affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal?</i> Respond with label only
Style: <i>Few Shot Learning with Split Definition</i> Using the following LIWC2015 labels for the following words as example: “fake” labeled as “affect” ... “principal” labeled as “drives” ... “mwah” labeled as “informal” For a word with the following meaning: “a word commonly used to describe an emotional state in which the person feels a sense of having no hope; usually during a deep depression.” in this example “As I lay awake, alone in my bed, I cannot help but become overwhelmed by this feeling of blah.” which label should be assigned to the word blah from the following “LIWC2015” labels: <i>affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal?</i> Respond with label only
Style: <i>Few Shot Learning with Definition as Context</i> Using the following LIWC2015 labels for the following words as example: “fake” labeled as “affect” ... “principal” labeled as “drives” ... “mwah” labeled as “informal” For a word with the following meaning: “a word commonly used to describe an emotional state in which the person feels a sense of having no hope; usually during a deep depression.” in this example “As I lay awake, alone in my bed, I cannot help but become overwhelmed by this feeling of blah.” which label should be assigned to the word blah from the following “LIWC2015” labels: <i>affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal?</i> Respond with label only

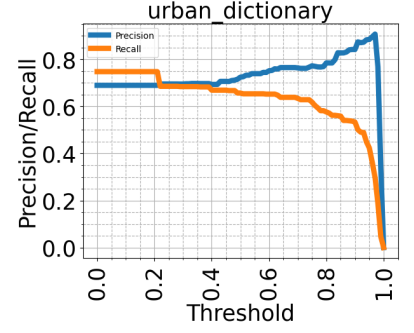
Table 20: Examples of different prompt style we experimented with. **Bold text** indicates a template used with every prompt. *Italic text* indicates lexicon dependant part of the template. The rest is the content of the definition.



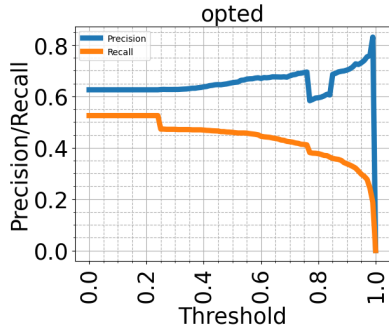
(a) Values with OPTED



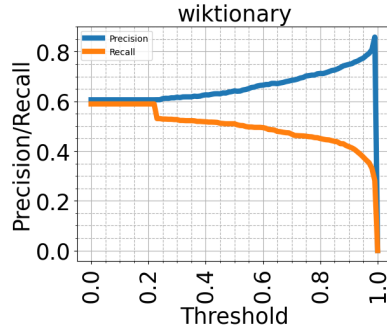
(b) Values with WK



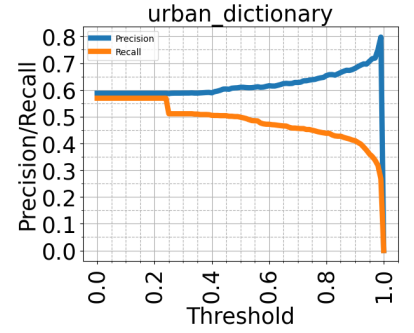
(c) Values with UD



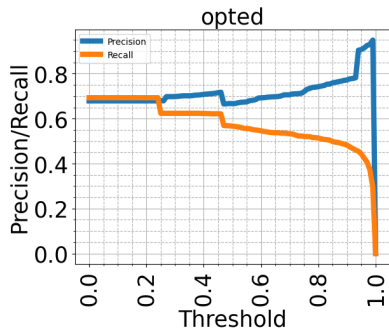
(d) LIWC2015 with OPTED



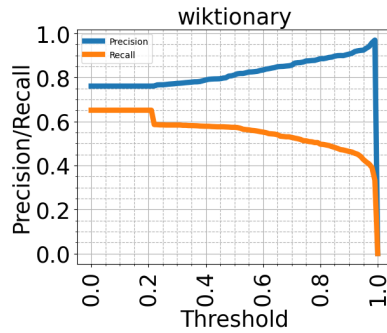
(e) LIWC2015 with WK



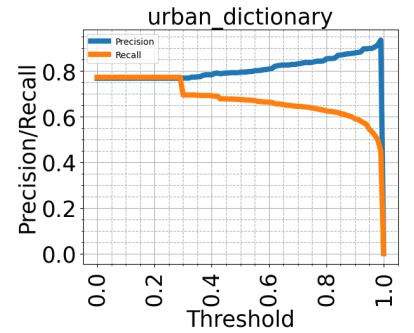
(f) LIWC2015 with UD



(g) LIWC22 with OPTED



(h) LIWC22 with WK



(i) LIWC22 with UD

Figure 2: Precision and Recall curves against model confidence for different lexicon-dictionary pairs.

Prompt Style	Precision	Recall	F-Score
ChatGPT 3.5 with zero shot learning, no lexicon name	0.24	0.27	0.20
ChatGPT 3.5 with zero shot learning, lexicon name	0.25	0.27	0.20
ChatGPT 3.5 with few shot learning, lexicon name	0.40	0.35	0.30
ChatGPT 4 with few shot learning, lexicon name	0.43	0.41	0.38

Table 21: Macro-average scores on the full test set for different prompts styles. Note that for all cases we prompt the model to return the label only.

	OPTED	WK	UD
Values	7.90	2.63	1.43
LIWC2015	2.51	2.73	1.28
LIWC22	3.49	7.37	4.16

Table 22: Terms with two or more definitions, the ratio of consistent labels to differing ones.

than 0.98 confidence. All definitions were related to sleeping in a car. The above are shown in Table 23.

To quantize how consistent or redundant definitions are within the same dictionary, we compute the ratio of terms with two or more definitions assigned the same label for all definitions and the number of terms assigned more than one. The ratios are shown in Table 22.

E Limitations

The proposed work only addressed categorical lexicons which are composed of term-labels pairs. This does not cover other types and structures of lexicons such as WordNet. Nevertheless, dictionaries can be beneficial in expanding such lexicons (using similarity for example) but it is out of this work’s scope. Another challenge dealing with categories with very limited number of entries. In that case, new candidate terms for these categories are also limited to none which does not allow us to analyse or compare the dictionary impact on expanding these terms. Our work also used the publicly and freely available dictionaries but not other dictionaries that would require subscription. While these were convenient for the current research purposes, the use of other dictionaries can provide more insight on evaluating Lexpansion. The presented Lexpansion and analysis only addressed labelling terms with a single category although the expanded lexicons do not have such a constraint. Our work also does not address possible biases that might exist in crowd sourced dictionaries such as in the case of UD which contains significant offensive stereotypes. We are also unable to generate new terms for lexicon categories that were never matched by the lexicon when annotating the dictionaries to generate the training and testing set. Finally, while our method can be applicable to different languages, this work was only addressed English.

Term	Lexicon	Dictionary	Label
emo	LIWC2015	UD	
	music bands depressed sometimes fun style unique 1) An emotional person. They are not depressed all the time and some are actually very happy at times. They do smile, they don't sit in a corner crying all day. Some are actually quite popular and laugh and joke around lots ...		affect
	a) short for the term emotional...in a musical sense b) music derived in the 80's... with such bands as Rites of Spring, Texas is the Reason, and more. c) can be used to describe a person who listens to emo, can relate to most of it and then cry because they can relate to it and not just because its emo. emo music is not punk. emo music usually contains lyrics which have a desperate side to them ...		affect
	music selling out corporate rock whoredom bland armchair rebels Bland, shallow corporatist pastiche of Punk that embodies none of the values or raw honest emotional integrity of Punk but continually dances on its grave none the less.. Strangely popular. I just saw Emo kings My Chemical Romance advertise Guitar Hero on Xbox Live Marketplace ...		affect
	Emo is genre of hardcore punk music started in the mid eighties by a band called Rites Of Spring. Their lead singer had been a big fan of the band Minor Threat, and after their breakup, had decided to start his own band. It was similar in the musical sense, however the lyrics were different ...		pconcern
	LIWC2015	WK	
	a young person who is considered to be over-emotional or stereotypically emo.		affect
	depressed. criticism drapes a black velvet cape across the puddle that interrupts the path to change, to be emo about it.		affect
	associated with youth subcultures embodying emotional sensitivity. the one thing everyone agrees on is that they've never encountered a band that claimed to be emo. trevor looks kind of emo, rail thin, dark hair, guyliner, wears black all the time.		affect
duh	LIWC22	UD	
	Duh means "No sh*t sherlock" and/or "Thank you captain obvious" Him: I'm 18 Me: DUH!		Conversation
	An expression that someone says when another person says something obvious or dumb. Person 1: Hey look I can see myself in the mirror! Person 2: duh.		Conversation
	replacement for the sarcastic retort, "No Kidding!" "Hey, the Sun is bright" "Duh!"		Conversation
	LIWC22	WK	
	Disdainful indication that something is obvious. it's hot in the desert. - well, duh!		Conversation
carcolepsy	Indication of mock stupidity. duhhh, I'm jasmine, I can't even tie my shoe laces right!		Conversation
	LIWC22	UD	
	sleep attack sleep creep sleep hole sleep monster sleepathon. a condition affecting buddies on a trip who fall asleep as soon as the car starts moving, providing no company or driving help. Joe slept the whole way here, I think he suffers from carcolepsy.		Physical
	narcolepsy sleep roadtrip dws drive Pronounced: Kar-ko-lep-see The inability to stay awake and alert when in a car, or any other thing that moves, such as trains, planes, and busses. The act of passing out while in a car regardless of passenger or driver status. Roadtrip? Count me in, but I can't drive. I have carcolepsy and we'd all die.		Physical
	sleep carcolepsy tired road trip nubman a condition characterized by brief attacks of deep sleep brought on by simply being in a car Dude, your car is so comfortable it gives me carcolepsy		Physical

Table 23: Example terms with multiple definitions across several dictionaries, and the lexicon categories that were predicted for these definitions.