Lexpansion: A Method for Automatic Lexicon Expansion using Dictionary Term Classification

Anonymous ACL submission

Abstract

Lexicons are useful resources for analysis and classification in NLP. Most Lexicons are static and require manual effort to curate and maintain. Regular updates are needed to maintain their relevance, especially with the evolution of language in terms of semantic shifts and neologisms. In this work, we explore the potential of supervised learning for expanding lexicons using dictionary entries. We study the effect of using dictionaries with varying properties such as size, labels, structure and method of curation. We use definitions as input features to a classification model to assign lexicon categories to terms, and analyse the proposed expansions in terms of estimated accuracy and coverage for each dictionary. Our results show promise for obtaining a considerable amount of new terms with high precision. We release our expanded lexicons, code, and pretrained models.

1 Introduction

Lexicons are important in NLP application 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 deep learning and LLMs, lexicons are useful to interpret results, build lightweight models or add emphasis on relevant features (Li et al., 2020).

However, lexicons are mostly static while language evolves over time. Various societal and demographic factors have been shown to affect language change (Labov, 2011; Milroy and Milroy, 2017), and these changes are amplified with the emergence of social media use both syntactically (Maity et al., 2016) and semantically (McGillivray et al., 2022). Only a very limited number of lexicons are updated to cope with these changes, e.g., LIWC (Pennebaker et al., 2001, 2007, 2015; Boyd et al., 2022) which is updated regularly, but only at 7-years intervals.

An alternative is using language resources such as dictionaries to expand lexicons (Bahgat et al., 2022). Unlike most lexicons, dictionaries are updated regularly and capture language changes (Nguyen et al., 2018). For example, in March 2023, Oxford English Dictionary (OED) added 700 new terms as "deepfake" and "groomzilla" while revising 1,400 entries[1]. Some dictionaries are more sensitive to changes than others. For example, in Urban Dictionary (UD)[2], a crowd-sourced dictionary, "selfie" was first defined in September 2009, while was added to OED in August 2013 (Nguyen et al., 2018). Crowd-sourced dictionaries has more slang that is used in social media and is useful in relevant applications (Wilson et al., 2020a).

In this work, we present a generic expansion method "Lexpansion" for lexicons by obtaining candidate new terms from dictionaries. We train a model using entries extracted from a dictionary that match existing lexicon terms. Input features are drawn from the dictionary definition to predict the lexicon label. The models are evaluated on examples matched on held out terms from the lexicon. Our framework allows (1) training a model once and then extracting candidates from multiple batches of data without the need to retrain the model and (2) comparing the results of using different dictionary to expand the same lexicon. We explore the impact of different dictionary properties on this process to inform the selection of the best option for each lexicon. Our method can reduce the cost of updating lexicons and allow them to adapt when language shifts. Our code[3], models and expanded lexicons are available publicly.

2 Related Work

Lexicons are used widely for analysis in many sociolinguistics and computational social science stud-

[1] Mar 2023 update https://www.oed.com/information/updates/march-2023/

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

[3] https://anonymous.4open.science/r/lexpander-F9A7/

ies. This includes analysing political speeches (Liu and Lei, 2018), discussion about elections on social media posts (Tumasjan et al., 2010), hate speech (ElSherief et al., 2018), social hierarchies and cohorts (Kacewicz et al., 2014), mental health disorders (Coppersmith et al., 2014) and emotions and values (Bahgat et al., 2020). In addition, lexicons are effective when added as additional features for machine learning models (Biggiogera et al., 2021; Cutler et al., 2021). Lexicons can be used as seeds to expand on specific features relevant to a target task (Baziotis et al., 2018). In other cases, lexicons are augmented with model training in which features extracted using lexicons are fed into the different neural layers to obtain domain-specific information (Lei et al., 2018; Li et al., 2020).

More inclusive and accurate analysis requires up-to-date lexicons to capture newly used words or new usage of existing ones. Automating the update and expansion of lexicons has garnered interest given the cost of manually building lexicons.

A common approach to expand lexicons is through using synonyms from existing resources, usually WordNet (Miller, 1998), to add new terms to an existing smaller seed lexicon. For example, work in Badaro et al. (2018) expanded *DepecheMood* lexicon (Staiano and Guerini, 2014) using synonyms for matched words. The new lexicon, *EmoWordNet*, contained 1.8 times more terms than the source lexicon. Similarly, Shaikh et al. (2016) added the use of hyponyms along with synonyms. The expanded version of *DepecheMood* in that case was 5 times the original size. As WordNet is itself a manually curated resource, it is expected that the quality of included terms is high, which in turn means the added terms are less noisy. Nevertheless, WordNet suffers from less frequent updates same as other manually curated resources[4].

Some work used dictionaries for expansion. Bahgat et al. (2022) used Urban Dictionary to expand categories in LIWC2015. The terms were classified by using their corresponding meaning and example. However, that work is limited to a single specific dictionary and does not investigate the impact of some dictionary properties. For example, when expanding LIWC2015, we obtained even better results using Wiktionary instead of Urban Dictionary. We also obtained only slightly lower results using a much smaller dictionary; OPTED. Our work builds on Bahgat et al. (2022) but departs from it by using multiple dictionaries and lexicons with varying properties and analysing outcomes. In another case, Di Natale and Garcia (2024) used shared senses between terms to find new candidates. Shared senses are obtained from aligning terms from different languages using language-to-language dictionaries in order to find new terms. That is, when a term t_{l_1} from a source language l_1 is mapped to a term t_{l_2} from a destination language l_2, all terms in the source language l_1 that correspond to the term t_{l_2} in the destination language are considered as candidates for adding to the same lexicon category that the term t_{l_1} in the source language l_1 belongs to.

Other approaches rely on word embeddings. In a word embeddings space, terms are represented by vectors. The differences between vectors are smaller for terms that are semantically similar (Mikolov et al., 2013). That characteristic was exploited by Empath (Fast et al., 2016) to select candidate new terms for manual review from the embeddings space that are neighbouring the existing terms in LIWC lexicon (Pennebaker et al., 2015). The space was built using fictional text that is expected to contain more emotional content that is more relevant to LIWC, a psychometric lexicon. Similarly in Wilson et al. (2018), a set of seed terms representing selected personal values are used to identity candidate terms neighbouring those terms.

Other work by (Avancini et al., 2006) expands lexicons into new target domains. Features based on *tf/idf* vectors generated from relevant domain corpora were used with *ADABoost.MH*KR classifier to label new terms. The resulting lexicon had higher coverage compared to the source lexicon on a test corpus for target domains.

Staiano and Guerini (2014) expanded *DepecheMood* to create *DepecheMood++* automatically. Crawled text from news articles website were used where users are able to annotate each article with 8 different emotion labels. They construct a document-by-emotion and word-by-document matrices to estimate the labels for each word.

In the above approaches, whether based on word embeddings, a classifier or matrices, those need to be recomputed to add new content which is contrary to our method where we can reuse the same classifier to classify new content without rebuilding it.

[4]Latest update was 2006 as mentioned in https://wordnet.princeton.edu/download/current-version

3 Lexpansion Methodology

Our methodology leverages dictionary term definitions as inputs to a classifier that predicts the lexicon category that are terms belongs to. Lexicon expansion is carried out through a series of experiments, where each experiment uses one or more dictionaries to train corresponding models. By utilizing multiple models within an experiment, we can assess and compare the influence of each dictionary on the quality of expansions.

3.1 An Expansion Run

A run is defined as an experiment e to expand a lexicon l. Experiment configurations include selection of dictionary entries, selection of categories, sizes of training and testing sets and model parameters. An experiment uses a set of dictionaries D to run multiple expansions for a lexicon. For each dictionary $d_i \in D$ in experiment e results are stored to compare the use of different dictionaries.

3.2 Lexicons

Lexicons are a set of entries and their corresponding categories. An entry t^l in lexicon l can be assigned one category c^l or more. Lexicon terms are matched through exact matches or wild cards.

We expand three lexicons; Values, LIWC2015 and LIWC22. These lexicons are hierarchical but vary in size and complexity. Example terms are illustrated in Section A.

Values is a lexicon that captures personal values (Wilson et al., 2018). The authors created an initial hierarchy with seed terms. More candidate terms are proposed by selecting neighboring 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".

LIWC2015 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 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 sub-categories. 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.

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

Table 1: Number of terms in each lexicon.

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.

LIWC22 is the next iteration after LIWC2015. LIWC22 has a revised set of categories with a slightly different hierarchy. Some categories were added as *Culture*, renamed as *Life Style* from *Personal Concerns* or deleted altogether as *Relativity*. LIWC22 has also an increased coverage using nearly twice as many entries compared to LIWC2015 and more flexible wild cards appearing at any place in terms. Also, while LIWC2015 only contained unigrams, LIWC22 n-grams such as "civil unrest" or "not in the mood".

Similar to LIWC2015 case, we expand a subset of LIWC22 top categories of LIWC22. The categories selected are: "Culture", "Conversation", "Cognition", "Drives", "Social", "Physical", "Affect", "Lifestyle" and "Perception".

Table 1 compares some lexicon parameters.

3.3 Dictionaries

Dictionaries contain a list of terms and their definitions as well as other information. A term t_i^d belonging to a dictionary d can have multiple definitions which can be represented as a set $H_{t_i^d}$. A definition h 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 use 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 adding new *definitions* to already existing terms. Definitions are composed of meaning and example. 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 as "happy : something that no Gen Z'er experiences" for "happy" or humor as "what UD wants you to buy all the time" for "mug"[5]. UD users are allowed to up-vote or down-vote a definition. Votes help definitions to bubble up, but nonetheless some of the highly voted definitions are still noise. A common case is proper nouns which usually describe a person known to the author with votes being based on popularity rather than soundness. UD also has a significant amount of offensive content (Nguyen et al., 2018). UD contains slang or informal definitions and terms (Wu et al., 2018) that are commonly used on social media and is sensitive to new terms surfacing on social media (Wilson et al., 2020b) or news (Bahgat et al., 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 though, we use meanings and examples only. WK also has more content compared to formal dictionaries. But then compared to UD, WK has less number of terms and definitions as shown Table 2.

The Online Plain Text English Dictionary (*OPTED*)[6], 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 that are 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. Also 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 "innoxious" which is defined as "Free from crime; pure; innocent.".

[5]Referring to mug advertisements on UD website.

[6]OPTED 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/

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.

Dictionaries are processed to filter out noisy entries. In UD difference between likes and dislikes was used as a quality proxy to remove low quality definitions. Names and stop words were excluded.

Table 3 shows the number of matches for each lexicon per dictionary. Only terms in the selected categories or their descendants are included.

3.4 Selecting Train and Test Samples

Terms used for training and testing set are selected from the dictionaries. Starting with a single dictionary experiment, a training sample is created by matching a lexicon entry with a term in the dictionary. The definition would be the input and the corresponding label or labels would be the expected output. A definition is composed of a meaning and an example if one exists. A term can have multiple definitions in a single dictionary. In that case, each definition is considered a separate sample.

To create the test set, dictionary terms and their corresponding definitions are selected at random from the lexicon. If there is a quality metric provided, the definition with the highest quality is selected, else the first definition was selected under the assumption that this is the most common definition. We compute the distribution of lexicon labels over all the matched dictionary entries and match that distribution in the test set.

Next, samples for the training set are selected. Entries that match a test set term are excluded along with any corresponding samples. That includes terms that can be matched with wild cards. For dictionaries with no quality indicator, we assume that all definitions are of reasonable quality. In that case all definitions are used as training samples to get the maximum available. For dictionaries that provide a quality measure such as votes as in UD's

case, the top n definitions with highest quality are included in the training set. If one of the top n definitions has a defined quality indicator value below a certain threshold, that definitions can as well be excluded. That selection strategy provides a balance between quality and obtaining enough samples to train the model properly.

An experiment for a given lexicon can also be run using multiple dictionaries to compare results from each dictionary. To correctly compare, the content of the test set needs to match for the different dictionary runs. Thus, the test terms are selected such that they exist in the multiple dictionaries. Then the corresponding definition from each dictionary are added to the set. The selection of samples for the training set is done the same way as in the case of a single dictionary.

3.5 Model

We formulate lexicon expansion as a classification problem. A term is assigned a class representing a lexicon category based on the corresponding definition. 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.

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 threshold for confidence that would filter out terms assigned labels with less confidence that the threshold, would include fewer terms, but it is expected that the precision would improve. Depending on the desired outcome, a threshold can be selected that balances precision and recall.

3.7 Using Majority Vote

Precision can be improved by combining label assignments to a term from different dictionary models for the same lexicon. Two methods were used either by majority agreement or unanimous model agreement. Thus 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 asses the worthiness of the proposed method, we ran two baseline experiments. Each of the two baselines answers a different question. One is whether a trivial model achieves a comparable performance and the other is whether large language models (LLMs) are able to perform better the expansion task without fine tuning.

Max Category Count was used similar to Bahgat et al. (2022). Terms are assigned labels with most matches in the corresponding definition. That baseline could work as synonyms are expected to have the same label. To avoid any possible bias, the candidate term is removed from the definition if found. 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."[7]. The term "pie" was labeled as "bio" according to LIWC2015. The definition included the words "kidney", "dinner" and "desert" which are all "bio" terms while "social" label was mentioned once with the word "family" and "relativ" label mentioned once with the word "outer".

ChatGPT was also used. Large Language Models (LLMs) achieved disruptive performance in encoding knowledge and generating content. Output from such models is well formed sentences that takes into account consecutive prompts in chat scenarios. ChatGPT specifically achieved impressive results on a range of tasks (Bang et al., 2023; Laskar et al., 2023). Such a performance would bid the question of whether smaller models are still relevant or not. So we prompted ChatGPT to label terms with different LIWC labels given their definition. We experimented with a range of prompts to find the best. Details are in section B.

4.2 Train and Test Sets Splits

The test set for LIWC2015 and LIWC22 experiments used 1,000 samples while for values only 200 given its smaller term count. The training set counts are listed in Table 4. From a dictionary stand point, WK had the highest matches while OPTED had the lowest. From a lexicon stand point, LIWC22 had significantly more coverage. For LIWC2015 we opted not to use wild cards as we saw a significant reduction of performance due to noise. Values had the least matches.

[7] https://en.WK.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: Baselines performance on LIWC2015

For UD, to select training set samples 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.

5 Results and Discussion

We list results of our runs and discuss the candidate new terms. All reported metrics are macro-averaged as classes are imbalanced.

5.1 Results Compared to Baselines

We compare *Lexpansion* against *Max Count* and *ChatGPT with few-shots* to expand LIWC2015 with UD. The results on 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 was performing better than the general purpose built 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 too far behind when compared to models trained using WK which had the most data. UD was mostly better in Values and LIWC22 cases and slightly lagging behind WK (with less data) in case of LIWC2015.

5.3 Tuning Results and Using Voting

The results shown in Table 6 are when accepting labels with model confidence 0 or more. A higher precision can be obtained as high as above 0.8 by increasing the confidence threshold to 0.98. Figure 1 shows the curve for LIWC2015 with UD case. Figures for other cases can be found in section C.

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

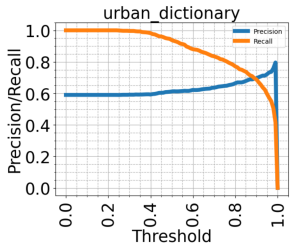


Figure 1: Precision recall curve for LIWC2015 vs UD

We have also combined the outcomes of models trained using different dictionaries such that only same labels assigned by two of the three models or all three are used. Results are shown in Table 7.

5.4 Expansion Counts

We run our models against all entries in each dictionary to generate a list of new candidate terms that can be used to expand existing lexicons. Table 8 shows the counts for each lexicon-dictionary pair that were assigned labels by the corresponding model with confidence more than 0.98. The table also shows the intersection between labels assigned from different dictionaries when only considering the label with highest confidence.

5.5 Agreement in the Same Dictionary

We look at the variations of labels assigned to terms that have two or more definitions within the same dictionary. Different definitions might convey close meanings for a term which should 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". "launder" was also labeled with "life" from

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 combining output of different dictionary models. "Vote" number of dictionary models agree on the same classification. "Count" shows samples that are in agreement and considered in the metrics.

	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*.

Values label with one definition is related to monetary transaction and the other with sanitation as both fall under the parent category "life" through two different sub categories; "wealth" and "health".

In other cases, definitions can be distinct conveying different meanings. In these cases the model should assign different labels. For example, the definitions for "abrade" illustrates a persons exhaustion in one definition versus a description of how a physical surface would feel in another. The corresponding LIWC2015 labels were "affect" versus "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 section D.

5.6 Agreement Across Different Dictionaries

We compare labels assigned by our models for the same term in different dictionaries for the same lexicon. Definitions in different dictionaries can have different or similar meanings. For example, Table 12 shows definitions for "analytics". Those definitions conveyed the same meaning and so term was labeled with LIWC22 "Cognition" for all three dictionaries. In the case of "acai berry", OPTED did not have an entry while WK had a proper definition versus 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 in WK case. For "cheesy", the models for the three dictionaries assigned the same label "Physical" as those definitions were dealing with food. But there were also definitions in WK and UD that were conveying a sentiment towards a person. In both cases, the term got "Affect" label which does match the context of 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 original LIWC22 lexicon label, while UD's does not. Table 10 shows more examples of terms that were either matched or were different across different dictionaries.

Lexicon	Term	Top label when using		
		UD	WK	OPTED
Values	enact	life	life	life
	crisis	life	social	children
	nonmated	-	social	-
	clingy	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: Example of terms added

We study terms with similar definitions and how their assigned consistent labels are. Considering only terms with two or more definitions UD had the highest number of definitions per term. But when calculating the average number of unique labels assigned to terms with two or more definitions, the difference between UD and the two dictionaries were small. These values are shown in Table 11.

	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: Terms with two or more definitions, average definitions/term versus average unique labels assigned.

6 Conclusion and Future Work

In this work, we presented our"Lexpansion" method to expand static lexicons using regularly-updated dictionaries. Terms in dictionaries that are in also lexicons are used to train models to classify

Term	Lexicon	Dict.	Label	Definition
semantics	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.
wheedle	LIWC22	OPTED	Social	To entice by soft words; to cajole; to flatter; to coax.
			Social	To grain, or get away, by flattery.
launder	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.
abrade	LIWC2015	WK	affect	to wear down or exhaust, as a person; irritate.
			percept	to cause the surface to become more rough.
fam	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?
gale	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.

and add new terms to the lexicon. We used our proposed method to expand three lexicons using three dictionaries. Each of the dictionaries vary in properties including the use of formal versus informal language, sensitivity to new terms, term count, number of different definitions per term and others. Our results show the effectiveness of Lexpansion method. Our approach was demonstrated to be significantly more effective than other baselines including simple voting and LLMs based. For future work, we aim to identify metrics to quantify parameters as sensitivity to language shifts, ambiguity, and coverage. We would also apply our method to multiple languages.

	Dict.	Definition	Label
analytics	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
acai berry	OPTED	-	-
	WK	a fruit that grows on the brazilian wild palmberry tree, euterpe olearacea, used for nutritive support as an antioxidant. it is similar in size to a grape.	Physical
	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
cheesy	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
	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
assailant	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.

7 Limitations

There are limitations for the proposed method. For one, the size of the dictionary as well as the size of lexicon that needs to be expanded has an impact on the number of data for training and testing examples. Consequently, the resulting model is less accurate and produces lower amount of new terms with high confidence. This also means in some cases some categories do not get new terms at all.

Another limitation is the availability of open dictionaries. While Wiktionary, Urban Dictionary and OPTED are publicly and freely available, other bigger dictionaries do not provide dumps that can be used conveniently for research purposes. Subscription may be required to avail the data. This was one of the main reasons we opted for OPTED. While OPTED is small in size, but it is freely available.

In the lexicons we expanded, words can belong to multiple category. But this not the case for the current work. We only assign one category per term even though some have multiple definitions and could belong to multiple categories.

Another consideration is that Urban Dictionary contains offensive stereotypes which may get included in our model and some of its suggestions to expand lexicons with.

Finally, while our method can be applicable to different languages, our currently presented work was only tested for English.

References

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):1–30.

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*, pages 86–93.

Mohamed Bahgat, Steve Wilson, and Walid Magdy. 2022. Liwc-ud: Classifying online slang terms into liwc categories. In *14th ACM Web Science Conference 2022*, pages 422–432.

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*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 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*, pages 245–255, New Orleans, Louisiana. Association for Computational Linguistics.

Jacopo Biggiogera, George Boateng, Peter Hilpert, Matthew Vowels, Guy Bodenmann, Mona Neysari, Fridtjof Nussbeck, and Tobias Kowatsch. 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*, pages 385–389.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Andrew D Cutler, Stephen W Carden, Hannah L Dorough, and Nicholas S Holtzman. 2021. Inferring grandiose narcissism from text: Liwc versus machine learning. *Journal of Language and Social Psychology*, 40(2):260–276.

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)*, pages 4171–4186.

Anna Di Natale and David Garcia. 2024. Lexpander: applying colexification networks to automated lexicon expansion. *Behavior Research Methods*, 56(2):952–967.

Shihab Elbagir and Jing Yang. 2019. Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 122, page 16.

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*, volume 12.

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*, pages 4647–4657. ACM.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*.

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):125–143.

William Labov. 2011. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 3. John Wiley & Sons.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *arXiv preprint arXiv:2305.18486*.

Zeyang Lei, Yujiu Yang, and Min Yang. 2018. Sentiment lexicon enhanced attention-based lstm for sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Wei Li, Luyao Zhu, Yong Shi, Kun Guo, and Erik Cambria. 2020. User reviews: Sentiment analysis using lexicon integrated two-channel cnn–lstm family models. *Applied Soft Computing*, 94:106435.

Dilin Liu and Lei Lei. 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.

Suman Kalyan Maity, Bhadreswar Ghuku, Abhishek Upmanyu, and Animesh Mukherjee. 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)*, pages 1681–1690. IEEE.

Barbara McGillivray, Malithi Alahapperuma, Jonathan Cook, Chiara Di Bonaventura, Albert Merono Penuela, Gareth Tyson, and Steven Wilson. 2022. Leveraging time-dependent lexical features for offensive language detection. In *Proceedings of the 1st Workshop of Ever Evolving NLP, EMNLP 2022*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

James Milroy and Lesley Milroy. 2017. *Varieties and Variation*, chapter 3. John Wiley & Sons, Ltd.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

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):172320.

J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and Booth R. 2007. The development and psychometric properties of liwc2007. pages 1–22.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

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.

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)*, pages 1127–1132.

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)*, pages 427–433.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 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, pages 178–185.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Steven R Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020a. Urban dictionary embeddings for slang nlp applications. ACL.

Steven R Wilson, Walid Magdy, Barbara McGillivray, and Gareth Tyson. 2020b. Analyzing temporal relationships between trending terms on twitter and urban dictionary activity. In *12th ACM Conference on Web Science*, pages 155–163.

Steven R Wilson, Yiting Shen, and Rada Mihalcea. 2018. Building and validating hierarchical lexicons with a case study on personal values. In *International Conference on Social Informatics*, pages 455–470. Springer.

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):839–852.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 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*, pages 1496–1496.

A Lexicons Category Examples

Tables 13, 14 and 15 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 Labeling 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.

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 13: 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 14: 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 15: Examples from selected categories in LIWC22. The "*" symbol denotes a wildcard.

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 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 format[8]. The content 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 16.

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 16.

To query ChatGPT API, we used *text-davinci-003*[9] and *gpt-4-1106-preview*[10]. 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 permutations for 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 17 show results from experimenting the different choices of prompts. The zero shot learning styles used was "Definitions as Context" from Table 16, 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.

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,

[8]Using the *completion* API for ChatGPT.

[9]https://platform.openai.com/docs/models/gpt-3-5

[10]https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

Style: *Split Definition*
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:** *affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal***?** **Respond with label only**
Style: *Definition as Context*
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:** *affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal***?** **Respond with label only**
Style: *Few Shot Learning with Split Definition*
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:** *affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal***?** **Respond with label only**
Style: *Few Shot Learning with Definition as Context*
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:** *affect, social, cogproc, percept, bio, drives, relativ, pconcern, informal***?** **Respond with label only**

Table 16: Examples for some of the 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.

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 17: 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.

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 the 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 Labeling Consistency with Multiple Definitions

Definitions in UD are commonly repeated or even redundant conveying similar meanings. For example, the term "emo" had 315 definitions labeled 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 definitions 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" which is only found in UD. That term had 5 definitions or total of 13 when including ones labeled with less than 0.98 confidence. All definitions were related to sleeping in a car. The above are shown in Table 18

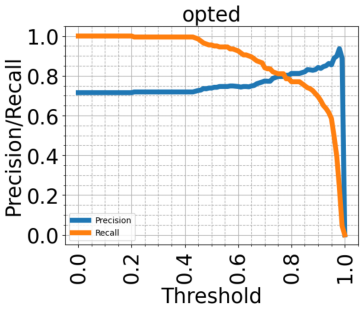
To quantize how consistent or redundant definitions are within the same dictionary, we compute the ratio of terms with two or more definitions assigned the same label for all definitions and the number of terms assigned more than one. The ratios are shown in Table 19.

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

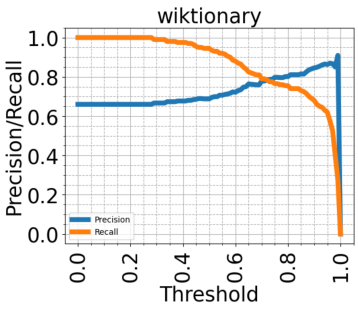
Table 19: Terms with two or more definitions, the ratio of consistent labels to differing ones.

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 acually 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 non of the values or raw honest emotional integrity of Punk but continually dances on its grave non 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
	Indication of mock stupidity. duhhh, I'm jasmine, I can't even tie my shoe laces right!		Conversation
carcolepsy	***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

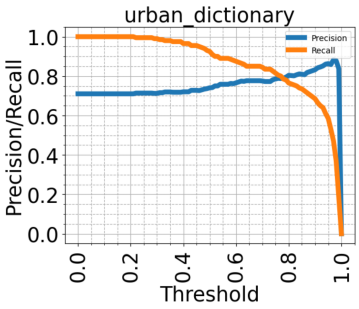
Table 18: Examples terms showing for each similar definitions corresponding to the same term in different dictionaries.



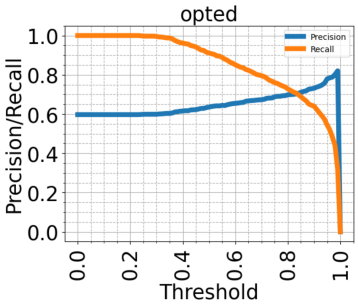
(a) Values with OPTED



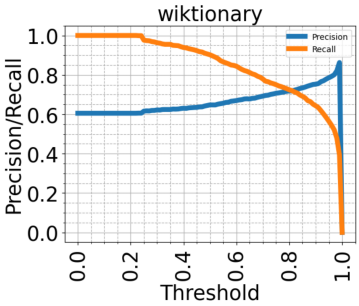
(b) Values with WK



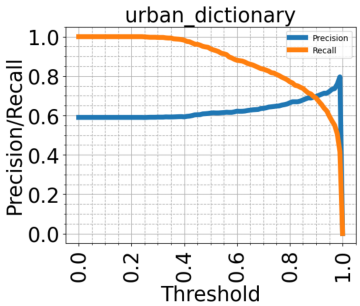
(c) Values with UD



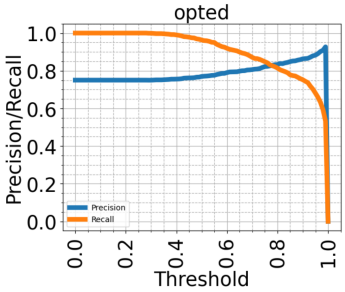
(d) LIWC2015 with OPTED



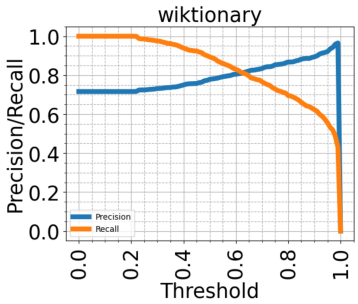
(e) LIWC2015 with WK



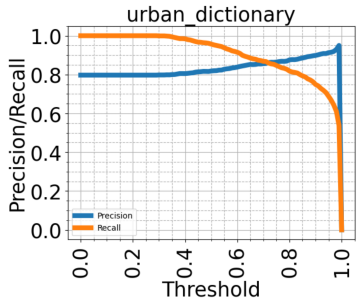
(f) LIWC2015 with UD



(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.