Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 61–69
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Sense and Sentiment

Francis Bond and Merrick Choo Yeu Herng

♠ Department of Asian Studies, Palacký University, Olomouc
◇ Computational Linguistics Lab, Nanyang Technological University
bond@ieee.org, merrick.emrys@gmail.com

Abstract

In this paper we examine existing sentiment lexicons and sense-based sentiment-tagged corpora to find out how sense and concept-based semantic relations effect sentiment scores (for polarity and valence). We show that some relations are good predictors of sentiment of related words: antonyms have similar valence and opposite polarity, synonyms similar valence and polarity, as do many derivational relations. We use this knowledge and existing resources to build a sentiment annotated wordnet of English, and show how it can be used to produce sentiment lexicons for other languages using the Open Multilingual Wordnet.

Keywords: sentiment, wordnet, derivation, inflection

1. Introduction

Sentiment analysis is the process of detecting, extracting, and classifying subjective information in a text (Lei and Liu, 2021). This can vary from a simple determination of positive or negative (polarity), to combining this with a strength (valency), analysing different scores for positive, negative and neutral, and potentially identifying information about emotions and arousal (affect). Lexicon-based and data-driven approaches are both widely used, and inadequacies in the existing analysis tools, such as lexicons and training datasets are considered important problems (Zunic et al., 2020).

The goal of this paper is to make a large, accurate, open sense-based sentiment lexicon for English. Currently we have some large sense-based lexicons such as ML-SentiCon (Cruz et al., 2014), but they are based on a small number of hand-seeded entries, and are not so accurate (Bond et al., 2019). Most other lexicons are based on either words or lemma and part-of-speech, which means firstly that differences in sentiment for different senses will be lost and secondly that it is difficult to map the scores accurately to different languages. We have built a sense-based sentiment system (**sensitive**) to go with this lexicon.[1]

We will first look at some existing sentiment resources for English and use them to shed some light on the nature of basic sentiment scores for words, senses and concepts: first lexicons § 2 and then corpora § 3. Then we will use these to build a new resource: the sentimental wordnet § 4, available under an open license (MIT). Finally, we conclude and discuss future work.

2. Lexicons

We assume that individual senses of words have a semantic orientation independent of context (**prior polarity**: Osgood, Suci, and Tannenbaum (1957)) and that this can be modelled by a numerical value. This has been assumed for words (Taboada et al., 2011), but as they note, it is better thought of for senses. For example, ***plot*** is negative meaning “plan secretly, usually something illegal” but neutral when meaning “devise the sequence of events in a literary work”, ***novel*** “new” is positive as an adjective, but neutral as a noun “book”[2].

There have been many papers comparing sentiment systems and lexicons, we use them to guide our choice of which resources to consider. Ribeiro et al. (2016) compare twenty four sentiment analysis systems. For those with sentiment lexicons, they convert them to a common format and use the Vader system (Hutto and Gilbert, 2014) to compare them (see Section 2.5 for a description of the system). They compared eight lexicons in this way and found VADER, AFINN, Opinion Lexicon, Sentiment 140 Lexicon and So-Cal to perform the best (Ribeiro et al., 2016, Table 8). Reagan et al. (2017) show that the best results came when lexicons’ sentiment scores are fine grained, rather than just positive or negative. We therefore decided not to report on the Opinion Lexicon which only shows polarity, not valence. We looked at AFINN (Nielsen, 2011) but excluded it from further discussion in this paper, even though it is a well built multilingual resource, as it is considerably smaller than VADER and contains less information.

In addition to these, we also look at three lexicons released since 2016: the WKW-SCI Sentiment Lexicon (Khoo and Johnkhan, 2018),[3] a recently released resource that encodes sentiment differentiated by part of speech, the Glasgow Norms (Scott et al., 2019) which also has some words rated by sense and labMT (Dodds et al., 2015) a large collection of sentiment lexica in multiple languages. In comparison to Ribeiro et al.

[1] https://github.com/bond-lab/sensitive

[2] Note that a purely word based system will not be able to distinguish different parts of speech, so will potentially include totally unrelated meanings

[3] `https://blogs.ntu.edu.sg/chriskhoo/2017/07/wkwsci-sentiment-lexicon-v1-1-available-for-download/`

(2016), our goal is not to rate the lexicons, but rather to see what they can teach us about sentiment, and how we can use that to produce a single improved resource. The sizes of the lexicons and some brief statistics are given in Table 1. All valences are normalized to between -1 and $+1$. The total score is the result of adding the scores for all entries in the lexicon. We can see that lexicons with more raters tend to have lower polarity scores on average, both for an individual word (we show *good*) for the maximum and minimum values: a single rater can be more extreme, but this is reduced when you average over many people. Lexicons based on an independent wordlist are generally positive on average, those based on words originally marked for affect are negative, although the difference is small. The outlier is the WKW lexicon which is based on an independent wordlist but quite strongly negative. We have no explanation for this. Note that even for a prototypical positive word, such as *good* there is a great deal of variation, excluding the machine-learned lexicon, the average is 0.58 with a range from 0.47 to 0.67. In the next sections, we introduce each resource as well as two wordnet based lexicons.

2.1. Glasgow Norms

The Glasgow Norms are a set of normative ratings for 5,553 English words on nine psycholinguistic dimensions: arousal, valence, dominance, concreteness, imageability, familiarity, age of acquisition, semantic size, and gender association. The Glasgow Norms are unique in several respects. The lexicon itself is relatively large, while simultaneously providing norms across a substantial number of lexical dimensions although we only look at valence here. The lexicon contains 379 ambiguous words that are presented either alone (e.g., toast) or with information that selects an alternative sense (e.g., toast (bread), toast (speech)). For valence, words were rated on a nine point scale, with an average of 32.98 raters/word (ranging from 15-70) (Scott et al., 2019, Table 3). The paper and lexicon are made available under a Creative Commons Attribution 4.0 International License (CC BY).

2.2. labMT Lexicon

Dodds et al. (2015) created 10,000 word lexicons for 10 languages: English, Spanish, French, German, Brazilian Portuguese, Korean, Mandarin Chinese (Simplified), Russian, Indonesian, and Arabic.

Words were selected according to frequency in corpora, with the most frequent 10,000 annotated on a scale of 1 (most negative) to 9 (most positive) with 50 ratings per word. Words were not lemmatized, so inflected forms may have different values.

2.3. Sentiment 140 Lexicon

The sentiment 140 lexicon is made by learning sentiment from a large twitter corpus. Tweets with positive emoticons are assumed to have positive, and those with negative emoticons negative (Go et al., 2009). We include it to show how different machine-learned lexicons are from hand-built ones.

2.4. So-Cal

The SO-CAL lexicon was made by a single researcher, and checked by a committee of three (Taboada et al., 2011), some cross POS-checking also took place, and the adjectives were evaluated compared to Mechanical Turk raters. It has lemmas and parts of speech as it is designed to be used with pos-tagged and parsed text. The data is released as CC-BY-NC-SA.

2.5. Vader

Vader (Hutto and Gilbert, 2014) is a sentiment analysis system based on a large hand-built lexicon and a few heuristics: valence is adjusted for use of punctuation, capitalisation, negation, intensification and contrastive conjunctions (*but*). Ribeiro et al. (2016) experimented by replacing just the sentiment lexicons in multiple systems and found that VADER performed best of all the lexicon-based systems. They therefore used it to test the different lexicons. The lexicon was given valence scores using mechanical turk, with the average of several raters taken (a wisdom-of-the-crowd approach). Words were given ratings based on surface form, with inflected forms included. A substantial list of emoticons was also included. The data was made available under an open licence (MIT).

2.6. WKW-SCI Sentiment Lexicon

The lexicon is based on the 12dicts common American English word lists compiled by Alan Beale[4] from twelve source dictionaries (Khoo and Johnkhan, 2018). Version 1.0 contains 29,729 words tagged for valence with parts-of-speech. In addition it has intensifiers (*greatly*), mitigators (*fewer*), maximizers (*entirely*), minimizers (*minimal*) and negators (*neither*). The lexicon comprises 3,187 positive words, 7,247 negative words and 19,295 neutral words. The data is released as CC-BY-NC-SA.

2.7. Comparison of the Lexicons

Table 3 shows how well the scores correlate with each other. We use Pearson correlation coefficient (ρ) to compare the lexicons: a value of 1 is a perfect correlation, 0 would be no correlation. The three lexicons build using many annotators (LabMT, GLAD and Vader) correlated well with each other, with ρ = 0.95–0.96. The two lexicons that also include POS (SOCAL and WKW) correlate with these three with ρ between 0.86 and 0.90. S140 is clearly an outlier, with very low correlation — it is learned from tweets with no claim to generality.

We also measured the correlation by part of speech for SOCAL and WKW. It is 0.86 for noun, adjectives and adverbs but only 0.82 for verbs.

[4] `http://wordlist.aspell.net/12dicts-readme/`

Name	Size	Positive	Negative	Min	Max	Valence	*good*	Misc	#	Licence
GLAS	5,553	3,362	2,135	-0.79	0.73	0.02	0.63	wrd	33	CC-BY
labMT	10,222	7,152	2,977	-0.93	0.88	0.09	0.55	wrd	50	CC-BY-NC-SA
S140	62,468	38,312	24,156	-1.00	1.00	0.07	0.16	ML	0	Research
SOCAL	6,091	2,477	3,611	-1.00	1.00	-0.07	0.60	lex+pos	1	CC-BY-NC-SA
VADER	7,506	3,337	4,169	-0.97	0.85	-0.05	0.47	wrd	10	MIT
WKW	28,955	3,103	7,095	-1.00	1.00	-0.18	0.67	lex+pos	3	CC-BY-NC-SA

Table 1: Basic Comparison
The table shows the number of entries, the number with positive and negative valence, the minimum and maximum valence, the total valence, the score for the word *good*, the type (wrd is word based, lex+pos is based on the lexeme + part of speech and ML is machine learned), the number of raters and the license.

Name	POS	Size	Positive	Negative	Valence
WKW	a	7,719	1,516	2,503	-0.11
WKW	n	13,312	728	2,508	-0.25
WKW	r	2,512	609	771	-0.06
WKW	v	6,189	333	1,461	-0.28
SOCAL	a	2,819	1,246	1,572	-0.04
SOCAL	n	1,539	539	1,000	-0.12
SOCAL	r	877	448	429	0.02
SOCAL	v	1,130	345	785	-0.15

Table 2: POS Comparison

Part of speech can be used a proxy for sense. For example, in WKW, ***sublime***$_a$ has a score of 1.0, while ***sublime***$_v$ has a score of 0. Presumably these match the different senses "worthy of adoration or reverence" and "change directly from a solid into a vapor". We investigated how often the sentiment of words differ when the POS differs. It turned out that very few words differ: for WKW only 579 out of 28,955 (1.6%) and for SOCAL 187 out of 6,091 (3.1%). If we consider absolute differences > 0.2 (for a scale from -1 to $+1$) then only 139 and 16 entries differ respectively, far fewer than 1%. If a word is positive or negative with one part-of-speech it is very likely to be so with another.

2.8. SentiWordNet

SentiWordNet (Esuli and Sebastiani, 2006) is the first sense-based sentiment lexicon. It annotates synsets from Princeton Wordnet (Fellbaum, 1998) with three numerical values in the range $\langle 0, 1\rangle$ placing the synset in a three dimensional polarity space. The dimensions describe "how objective, positive, and negative the terms contained in the synset are". As the three values must sum to one, there are only two degrees of freedom.

About 10% of the adjectives were manually annotated, each by 3-5 annotators and then the scores calculated through the the definitions and propagated through the network (Baccianella et al., 2010). In SentiWordNet 3.0, the automated annotation process starts with all the synsets which include 7 "paradigmatically positive" and 7 "paradigmatically negative" lemmas.[5] The initial seed is expanded with a random walk algorithm to generate a training set for a committee of classifiers which estimates the final polarity scores of synsets. In the end, SentiWordNet 3.0 added automatic sentiment annotation to all of Princeton WordNet 3.0.

2.9. ML-SentiCon

The method proposed in Baccianella et al. (2010) has become the motivation for further work on the development of word-level and sense-level sentiment lexicons. ML-SentiCon (Cruz et al., 2014) expands the idea presented in (Baccianella et al., 2010) by introducing additional sources of information such as WordNet-Affect (Strapparava and Valitutti, 2004) and General Inquirer (Stone et al., 1966) to improve the accuracy and coverage of initial polarity seed. The seed is expanded using the same general approach proposed in Baccianella et al. (2010). However, instead of a single score for each synset, individual scores for each sense are calculated, and then the final synset scores are calculated by averaging these.

3. Corpora

The second resource we looked at was the sentiment marked corpus NTU-MC (Tan and Bond, 2012; Bond et al., 2019). In this corpus, stories taken from the Sherlock Holmes canon by Arthur Conan Doyle are sense-tagged and then marked for sentiment. Annotation has proceeded in three phases. In phase 1, the stories *The Adventure of the Dancing Men* and *The Adventure of the Speckled Band* were annotated in Chinese, English and Japanese and the results compared across languages (Bond et al., 2016). All concepts (words that appear in

[5] good, nice, excellent, positive, fortunate, correct, superior; bad, nasty, poor, negative, unfortunate, wrong, inferior (Turney and Littman, 2003)

Name	labMT	S140	SOCAL	VADER	WKW
GLAS	0.95	0.48	0.88	0.96	0.81
labMT		0.41	0.86	0.95	0.73
S140			0.51	0.57	0.35
SOCAL				0.90	0.86
VADER					0.90

Table 3: Correlation of Scores
Calculated using Pearson's ρ for all entries that appear in each pair of lexicons

wordnet) that, in context, clearly show positive or negative sentiment are annotated. For example, in (1), the appropriate senses of ***false***$_a$ and ***villain***$_n$ are annotated as -0.34 and -0.64 respectively.

1. If we make one false move the villain may escape us yet . *The Hound of the Baskervilles*

Operators such as ***very*** and ***not*** were not tagged. Concepts can be multiword expressions, for example *give rise* "produce" or *kuchi-wo hiraku* "speak". The corpus also contains sentiment annotation of larger chunks and the full sentences, but we will not use that here.

Annotation was done using **IMI** — A Multilingual Semantic Annotation Environment (Bond et al., 2015), extended to allow for the annotation of sentiment at concept and chunk level. We use a continuous scale for tagging sentiment, with scores from -100 to 100. The tagging tool (IMI) splits these into seven values by default (-95, -64, -34, 0, 34, 64, 95), and there are keyboard shortcuts to select these values. Annotators can select different, more fine-grained values if they desire.

The annotators were told to tag using several evaluative adjectives as guidelines, shown in Table 4. The table also shows new examples from the corpus after annotation.

Score	Examples		Corpus Examples
95	fantastic	very good	perfect, splendidly
64	good	good	soothing, pleasure
34	ok	sort of good	easy, interesting
0	beige	neutral	puff
-34	poorly	a bit bad	rumour, cripple
-64	bad	bad	hideous, death
-95	awful	very bad	deadly, horror-stricken

Table 4: Annotator guidelines for sentiment scores

In Phase 1, each of the three texts was annotated by a single native speaker for that language, then the different languages were compared, major differences discussed and, where appropriate, retagged. If they were not sure whether the text segment shows sentiment or not, annotators were instructed to leave it neutral (0). The final correlation across languages was between 0.73 and 0.77 (measured using Pearson's ρ).

Bond et al. (2019) then compared the values from this corpus with the sense annotation of the Polish wordnet (Piasecki et al., 2009; Zasko-Zielinska et al., 2015). Again, they found a reasonable cross-lingual correlation, in this case 0.65. They examined the major errors (that is when the two resources differed in polarity) and found 14 instances, all of which were errors in annotation that have since been corrected.

Finally they compared the two resources with the Micro-WNOp lexicon[6] — a **sense-tagged** sentiment lexicon used to evaluate SentiWordNet and build ML-SentiCon. The original version consists of 1,105 Wordnet synsets chosen from the General Inquirer lexicon (Stone et al., 1966) and annotated by 1–3 annotators. They found that human inter-annotator correlation was 0.88. We used version 3.0, which only has 1,054 entries, of which 457 are non-zero.[7] Correlation with the NTU-MC corpus was 0.75 (with only 130 entries found) and SentiWordNet was a much lower 0.63 (but for all entries). Micro-WNOp was used to build ML-SentiCon so cannot be evaluated with it.

Since these results, we have added sentiment annotation to several new stories in the NTU-MC, shown in Table 5. All were sense tagged by students, as described in Bond et al. (2021). In Phase 2, the sense annotation was done by multiple students, but the sentiment annotation was done by one RA (a student who did well in the course). This covered two new stories (*The Red-headed League* and *A Scandal in Bohemia*) and half of the novel *The Hound of the Baskervilles*. In Phase 3 (with an improved tool), both sense and sentiment were annotated by multiple student annotators, with a round of comparison and harmonization. One of the annotators is an automatic annotator based on the existing sentiment scores, so the raters have access to this information. This covered two more stories (*The Adventure of the Final Problem* and *The Adventure of the Naval Treaty*) and the rest of *The Hound of the Baskervilles*.

We summarise the corpus in Table 5. We can clearly see the wisdom of the crowds in effect: Phase 1 has effectively three annotators but for different languages (0.78), Phase 2 only one (0.68) and Phase 3 two to four with an extensive discussion of differences, this gets the highest score (0.80). The overall correlation with Micro-WNOp is still high, at $\rho = 0.75$ and the coverage has increased threefold. Note that the results for Phase 1 are slightly (0.78 rather than 0.75) better than those reported in Bond et al. (2019), as errors in an-

[6] `http://www-3.unipv.it/wnop/`

[7] Seven entries had bad synset identifiers, we fixed them and pushed the changes upstream.

Corpus	Sentences	Words	Concepts	Distinct	Pos.	Neg.	D Pos	D Neg.	ρ	Overlap
Phase1	1,199	23,093	13,077	3,504	983	1,244	618	660	0.78	130
Phase2	3,021	54,698	30,287	6,046	805	869	415	520	0.68	225
Phase3	3,250	60,702	33,407	6,102	1,556	4,710	783	959	0.80	240
NTUMC	7,470	138,493	76,771	9,741	3,344	6,823	1,340	1,631	0.75	339

Table 5: Corpus summary
Neg and Pos show the number of concepts with positive and negative sentiment (above a threshold of 0.05) and D Pos and D Neg show these for distinct concepts. ρ is the agreement with Micro-WNOp, and Overlap is the number of entries in Micro-WNOp annotated in the corpus.

notation have been corrected after the comparison with the Polish wordnet.

Overall, we finish with a sense based lexicon for over 24,000 concepts, with non-zero scores for just under 3,000. Because the annotation is based on a corpus, frequent concepts will be covered first, which is important, as Reagan et al. (2017) have shown that if you only have sentiment scores for a limited number of words, high frequency ones are most useful in calculating document sentiment.

3.1. An Analysis of the Effect of Semantic Relations

Despite the use of semantic relations in creating resources such as SentiWordNet and ML-SentiCon, there has been no empirical analysis of the effects of semantic relations on sentiment score. Zasko-Zielinska et al. (2015) note that synonyms (senses within the same synset) can have varied sentiment scores, even of different polarity, but that this is very rare. They also assume that antonyms will have the opposite polarity but equal score, but do not test this.

We found some examples of synonyms with different polarity in the Phase 3 results, but these were mainly errors in annotation. For example ***white*** "being of the achromatic color of maximum lightness; having little or no hue owing to reflection of almost all incident light" was given a negative score even though it appears to be neutral in meaning. On examination of the sentence in which appeared, we judge that "anemic looking from illness or emotion" should have been the correct tag: *His dark eyes, glaring out of the white mask of his face, were full of horror and astonishment as he gazed from Sir Henry to me.*[8] There was only one place where we thought a single synset was truly ambiguous, and that was ***pride*** which only has a single meaning in wordnet. However, many lexicons distinguish more: e.g. from wiktionary "Feeling honoured (by something); feeling happy or satisfied about an event or fact; gratified" (positive) vs "Having too high an opinion of oneself; arrogant, supercilious " (negative). We think it would be better to distinguish between these two senses.

	Synsets	Score	Lemmas	Score
All	9,416	-0.021	11,153	-0.021
Non-Zero	2,671	-0.073	2,989	-0.079
Positive	1,171	+0.289	1,296	+0.305
Negative	1,500	-0.355	1,693	-0.373

Table 6: Corpus-based Sentiment Wordnet Summary

Relation	All	Score	Non-Zero	Score
similar	833	+0.109	450	+0.202
hyponym	851	+0.075	312	+0.206
holo location	0	+nan	0	+nan
holo member	24	+0.007	2	+0.089
holo part	160	+0.013	12	+0.171
holo portion	0	+nan	0	+nan
holo substance	8	+0.021	1	+0.170
entails	58	+0.057	23	+0.143
causes	24	+0.093	9	+0.249

Table 7: Concept based relations

The NTU-MC contains new synsets not yet merged with the Open English Wordnet (McCrae et al., 2020), in this section we will only consider those part of the Princeton English Wordnet (v3.0) (Fellbaum, 1998). This means we have slightly fewer concepts (synsets), as shown in Table 6. Overall, the synsets are slightly negative, and only 28% of all synsets in the corpus are have non-zero polarity.

When we look at the concept level relations, in Table 7,[9] we see that overall, related concepts are close in sentiment. Looking across all relations is deceptive as most synsets have zero sentiment. We shall therefore discuss the numbers for non-zero synsets (that is those where one or both related sysnets have a non-zero value). The relations with closest sentiment value are `holonym-member`, `entails`, `holonym-substance` and `holonym-part`, although all with low numbers of examples. Surprisingly, `similar` and `hyponym` are almost the same. We see examples of hyponyms having a very different value from their hypernym, such as ***love*** 0.64 and ***hate*** -0.95 which are both hyponyms of ***emotion*** 0.0. If the hypnonym relations were typed in more detail, so we could tell the difference between an umbrella term (such as ***emotion***) and its children, and a term and a more specific term such as ***ardor*** (a hyponym

[8] From *The Hound of the Baskervilles.*

[9] These relations, and the sense-level relations are described in detail at `https://globalwordnet.github.io/gwadoc/`

of ***love***), then we could expect an even closer score.

Relation	All	Score	Non-Zero	Score
synonym	1,408	+0.069	551	+0.184
antonym	249	+0.217	116	+0.467
ant opposite	249	+0.081	116	+0.175
also	1	+0.000	0	+nan
derivation	1,293	+0.071	507	+0.180
pertainym	183	+0.110	115	+0.176

Table 8: Sense based relations

For the sense level relations, shown in Table 8, the scores are even closer. For `synonym`, we measure the average distance of all senses in the same synset. For `antonym`, we measure both the difference between related words, but also the difference when one has its polarity reversed (equal to the sum of their scores: `ant opposite`). So for ***love*** 0.64 and ***hate*** -0.95, which are related by antonymy, we compare 0.64 to 0.95 for a difference of 0.31 rather than 1.59. For all the sense relations for which we have sufficient evidence, it is clear that the sense relations are closely related.

In the next section, we will take advantage of these relations to expand the coverage of the sense-based **Sentimental Wordnet**.

4. The Sentimental Wordnet

In this section we describe how we build the **Sentimental Wordnet**. We take as its core the annotations derived from the NTU-MC.

In order to improve the coverage, we will add senses in three ways. The first two take advantage of the existing work on sentiment lexicons: (i) if a word is **monosemous** in wordnet and had a sentiment score in any of the three word based sentiment lexicons, we will give the lemma a score based on that. (ii) if a word plus part-of-speech combination is monosemous in wordnet and it has a sentiment score in either of the lemma+pos based lexicons, we will give the lemma a score based on that. Finally, (iii) we will use the semantic sense-links in wordnet to propagate scores: lemmas with no score will be given the same score as their synonyms, then lemmas linked by `derivation` and `pertainym`[10] will be given the score of the linked lemma, and those linked by `antonym` will be given the negative of its score.

4.1. Linking Monosemous Words

For each word in word-based lexicon, we look it up in wordnet, and if it only has a single sense, then the score for the word is assigned to the sense. For lexicons with both words and parts-of-speech, we look up the word and part-of-speech combination in wordnet, and if that has only a single sense, then it is assigned.

For example, the word ***damnable*** appears with only one sense “deserving a curse” and this is assigned the value -0.425. The word ***perk*** has only one sense as a noun “an incidental benefit”, with value 0.2, and one sense as a verb “gain or regain energy”, with value 0.0.

The sense-based lexicons made with these methods are compared with Micro-WNOp and the results shown in Table 9. In addition to the individual lexicons, we show the results of combining all the word based lexicons (WRD), all the lex+pos based lexicons (POS) and all of them together (LEX). We also show again the values of the corpus-based lexicon (NTUMC) as a comparison, then finally the result of combining them all (ALL).

Method	Size	≠ 0	ρ	Cover	≠ 0
VADER	1,700	1,700	0.95	40	39
GLAS	842	836	0.87	35	24
labMT	1,545	1,535	0.77	36	17
WRD	3,550	3,535	0.88	79	55
SOCAL	2,078	2,078	0.84	85	78
WKW	14,668	5,134	0.87	198	140
POS	15,002	5,744	0.85	207	148
LEX	16,499	8,179	0.86	217	153
LEX P	40,477	21,964	0.85	366	243
LEX P^2	52,740	29,608	0.84	476	321
LEX P^3	60,576	35,603	0.83	561	374
LEX P^4	65,226	39,630	0.82	612	401
LEX P^5	67,880	42,367	0.82	634	415
LEX P^6	69,394	44,329	0.81	648	423
LEX P^7	70,224	45,688	0.81	656	430
LEX P^8	70,719	46,596	0.81	659	430
LEX P^9	71,010	47,251	0.81	659	430
LEX P^{10}	71,161	47,661	0.81	662	432
LEX P^{11}	71,242	47,955	0.81	662	432
LEX P^{12}	71,297	48,156	0.81	662	432
NTUMC	11,154	2,989	0.75	339	200
ALL	26,325	10,793	0.81	471	296
ALL P	57,987	27,643	0.82	614	378
ALL P^2	69,788	35,683	0.82	695	431
ALL P^3	75,177	40,967	0.83	742	457
ALL P^4	77,546	44,125	0.82	758	464
ALL P^5	78,483	46,149	0.82	765	466
ALL P^6	78,905	47,454	0.82	767	467
ALL P^7	79,075	48,225	0.82	768	467
ALL P^8	79,131	48,772	0.82	768	467
ALL P^9	79,150	49,092	0.82	768	467
ALL P^{10}	79,155	49,343	0.82	768	467
ALL P^{11}	79,155	49,466	0.82	768	467

Table 9: Comparing Sense-based Lexicons
Comparison to Micro-WNOp. Size is the number of senses in the lexicon ≠ 0 is the number with non-zero sentiment; ρ is the correlation with Micro-WNOp, and cover is the number of entries in Micro-WNOp found in the lexicon (1,054 possible, with 457 non-zero)

The results show that the word-based lexicons correlate well with Micro-WNOp (0.77–0.95), but have very poor coverage (as only monosemous words are in-

[10] This is a relation between two senses where one is closely related to the other even though they are not the same part of speech: such as ***slow*** and ***slowly*** or ***moon*** and ***lunar*** (`https://globalwordnet.github.io/gwadoc/#pertainym`).

cluded). Surprisingly, despite the fact that they are all large, they do not overlap so much: combining them together in WRD doubles the coverage of any one lexicon, still with a high ρ of 0.88. Even then, they only include 79 out of the 1,054 test synsets. The lex+pos lexicons have much greater coverage. This is partly due to the fact that there are more monosemous entries if you also consider part-of-speech, so SOC has more coverage than labMT, even though it is smaller. More importantly, WKW also includes words with zero sentiment, and this makes its coverage much higher (207), with a high correlation of 0.87. This is close to the correlation of the WRD lexicons, which were built with many more annotators. Finally, combining them all (LEX) gives the best coverage (217), along with a ρ of 0.86.

The corpus based approach (NTUMC) starts off with better coverage (339 out of 1,054). Although it has fewer entries than the monosemous approaches, it does better with high frequency words. Combining them together (ALL) gives a decent lexicon, with almost 50% coverage (471) and a high correlation with the gold standard. All the correlations are higher than SentiWordNet (at 0.63), although it has the advantage of full coverage. This lexicon can still be considered handbuilt, although we expect it to have gaps in its coverage of medium frequency words.

4.2. Propagating through Semantic Links

In order to increase the size of the lexicon, while still keeping the quality high, we take advantage of the semantic links. Our **propagate method** goes through all lemmas in a lexicon and looks for lemmas linked by `derivation`, `pertainym` and `antonym` in wordnet. We then look at all lemmas in a given synset (`synonyms`) and give unlabelled senses the average of the scores of labelled senses. For example, ALL contains ***agreeability*** (0.47), ***agreeably*** (0.37) and ***disagreeable*** (-0.64), but not ***agreeable***! After propagation, ***agreeable*** "conforming to your own liking or feelings or nature" is given a value of (0.49), while ***agreeable*** "in keeping" is given a value of (0.0) in ALL P. The sense scores were consistently closer than the concept level scores (Section 3.1) so we only use them for propagation.

The approach is similar to that of Agerri and García-Serrano (2010), although they start from a smaller seed (just the ***quality*** synset), only make a binary decision (positive or negative) and use more relations (also `hyperonymy`, `hyponymy` and `cause`). Unfortunately, we could not find their lexicon online to compare with.

We show the results in Table 9, with propagation marked with P, and each level beyond one with a superscript: so P^{10} means that the propagation step has taken place 10 times.

Interestingly, for the lexicon, repeated propagation increases the coverage while gradually reducing the correlation. On the other hand, for the corpus+lexicon, the propagation slightly increases the correlation at first while repeated iterations increase the coverage with no loss in accuracy. Eventually, from both starts, the coverage saturates at over 70,000 senses (out of a total possible of 206,978 senses in Princeton wordnet 3.0), with a correlation of 0.83 (slightly below the human correlation of 0.88). Beyond 10 iterations for the lexicon seed and 7 for the combined seed, there is no appreciable increase in coverage. The addition of the corpus does better both in terms of coverage (an extra $\approx 8,000$ senses) and accuracy (correlation 0.82 rather than 0.81).

We are confident that this final lexicon is the best sense-based sentiment lexicon currently available. It is easy to update: if new words are added to wordnet, linked with semantic links, we may be able to propagate to them. If new corpora are annotated, we can add them.

The lexicon has around 79,000 senses marked for sentiment, with over 49,000 senses marked with non-zero sentiment. The coverage of sentiment bearing expressions is higher than any of the lexicons discussed in Section 2.

As the senses are linked through the Open Multilingual Wordnet (Bond and Foster, 2013), it can easily be used to make high quality lexicons in other languages. As discussed in Section 3, cross language correlation is around 0.75. Translation of words, either manual or automatic, cannot fail to add extra ambiguity, however sense-based translation preserves the meaning.

We should also note that, by virtue of being connected to a wordnet, the lexicon also contains many multi-word expressions. Reagan et al. (2017) point out the importance of these when investigating google books. If you treat the ***Great War*** or the ***Great Depression*** as two separate words, then they get a positive boost because of the presence of ***great***!

5. Conclusion and Future Work

We have created a new sense-based sentiment lexicon, based on wordnet. It has sentiment values for around 79,000 senses of which over half have some sentiment. Correlation with the gold standard is 0.83 (Pearson's ρ) slightly below human (at 0.88). The data and code are released at `https://github.com/bond-lab/sentimental` under an open license (MIT).

There are several areas in which we will continue this research.

- First, we will do one more round of quality control over the annotations in the NTUMC. We will compare all annotated words with the values predicted in the sentimental wordnet, and hand-check all those with a difference of greater than 0.5.

- We would like to investigate propagation using the hyponym hierarchy further. If we first check that all known children have the same polarity, for example, it may be appropriate to transfer the scores. This would greatly increase the coverage.

- We would also like to tag words in the example sentences of the updated Princeton WordNet Gloss Corpus (Rademaker et al., 2019). This is the approach taken by the Polish wordnet, and helps make sure all concepts have been considered.

- We have added many new words to our local version of the English wordnet, some of which have sentiment, we will continue submitting these upstream to the Open English Wordnet (McCrae et al., 2019). Adding sentiment for new words and concepts is important, we hope that the release of this resource will encourage even more work on high-quality lexical resources. In particular, we have added high freqeuncy sentiment-bearing words from VADER, which should improve the sentiment coverage considerably.

5.1. Acknowledgements

We would like to thank all the students involved in the tagging, from HG2002 (2011–2015, 2018–2021) and HG8011 (2016, 2018, 2019, 2021), especially Vigneswari Thirumoolan who did the Phase 2 tagging. This work was supported from the European Regional Development Fund Project "Sinophone Borderlands – Interaction at the Edges", CZ.02.1.01/0.0/0.0/16_019/0000791.

6. Bibliographical References

Agerri, R. and García-Serrano, A. (2010). Q-WordNet: Extracting polarity from WordNet senses. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362, Sofia.

Bond, F., Morgado da Costa, L., and Lê, T. A. (2015). IMI — a multilingual semantic annotation environment. In *ACL-2015 System Demonstrations*.

Bond, F., Ohkuma, T., Morgado da Costa, L., Miura, Y., Chen, R., Kuribayashi, T., and Wang, W. (2016). A multilingual sentiment corpus for Chinese, English and Japanese. In *6th Emotion and Sentiment Analysis Workshop (at LREC 2016)*, Portorož.

Bond, F., Janz, A., and Piasecki, M. (2019). A comparison of sense-level sentiment scores. In *Proceedings of the 11th Global Wordnet Conference (GWC 2019)*.

Bond, F., Devadason, A., Teo, M. R. L., and da Costa, L. M. (2021). Teaching through tagging — interactive lexical semantics. In *Proceedings of the 11th Global Wordnet Conference*, pages 273–283, University of South Africa (UNISA), January. Global Wordnet Association.

Cruz, F. L., Troyano, J. A., Pontes, B., and Javier Ortega, F. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Mitchell, L., Harris, K. D., Kloumann, I. M., Bagrow, J. P., Megerdoomian, K., McMahon, M. T., Tivnan, B. F., and Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of 5th Conference on Language Resources and Evaluation LREC 2006*, pages 417–422.

Christine Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision.

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eytan Adar, et al., editors, *ICWSM*. The AAAI Press.

Khoo, C. S. and Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.

Lei, L. and Liu, D. (2021). *Conducting Sentiment Analysis*. Elements in Corpus Linguistics. Cambridge University Press.

McCrae, J. P., Rademaker, A., Bond, F., Rudnicka, E., and Fellbaum, C. (2019). English WordNet 2019 — an open-source wordnet for English. In *Proceedings of the 11th Global Wordnet Conference (GWC 2019)*.

McCrae, J. P., Rudnicka, E., and Bond, F. (2020). English wordnet: A new open-source wordnet for English. Technical report, K Lexical News.

Nielsen, F. . (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages.CoRR*, pages 93–98.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up*. Wroclaw University of Technology Press. (ISBN 978-83-7493-476-3).

Rademaker, A., Cuconato, B., Cid, A., Tessarollo, A., and Andrade, H. (2019). Completing the Princeton annotated gloss corpus project. In *Proceedings of the 10th Global Wordnet Conference*, pages 378–386, Wroclaw, Poland, July. Global Wordnet Association.

Reagan, A. J., Danforth, C. M., Tivnan, B., Williams, J. R., and Dodds, P. S. (2017). Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Science*, 6(1):28, Oct.

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., and Benevenuto, F. (2016). Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):23, Jul.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., and Sereno, S. C. (2019). The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51(3):1258–1270, Jun.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Strapparava, C. and Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, June.

Tan, L. and Bond, F. (2012). Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4):161–174.

Turney, P. D. and Littman, M. L. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21(4):315–346.

Zasko-Zielinska, M., Piasecki, M., and Szpakowicz, S. (2015). A large wordnet-based sentiment lexicon for polish. In *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, pages 721–730.

Zunic, A., Corcoran, P., and Spasic, I. (2020). Sentiment analysis in health and well-being: Systematic review. *JMIR Med Inform*, 8(1):e16023, Jan.