Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

sentibank: A Unified Resource of Sentiment Lexicons and Dictionaries

Nick Oh

socius, London
nick.sh.oh@socius.org

Abstract

Sentiment analysis is critical across computational social science domains, but faces challenges in interpretability. Rule-based methods relying on expert lexicons enable transparency, yet applying them is hindered by resource fragmentation and lack of validation. This paper introduces sentibank, a large-scale unified database consolidating 15 original sentiment dictionaries and 43 preprocessed dictionaries, spanning 7 genres and 6 domains.

Introduction

Sentiment Analysis (SA), the automated process of identifying and extracting subjective information like opinions, emotions, and attitudes from text data, has become an increasingly critical technique across social science domains: ranging from Policy Making to Business Analytics, and from Social/Behavioural Analytics to Finance (Jawale and Sawarkar 2020; Raheman et al. 2022; Al-Qablan et al. 2023; Fioroni et al. 2023; Venkit et al. 2023). While deep learning models have excelled in achieving high accuracy, often surpassing simpler lexicon models in SA tasks (Al-Qablan et al. 2023), their inherently opaque nature poses challenges for applications in high-stakes domains like government policy making or mental health diagnosis, where transparent and interpretable decision-making is crucial (Rudin 2019; Jawale and Sawarkar 2020). Recognising the continued importance of rule-based SA, particularly in computational social science fields where interpretability is paramount, improving rule-based SA remains vital.

Rule-based SA relies on expert-curated lexicons containing words with pre-assigned sentiment scores, as human expertise is required to accurately annotate the sentiment of words across contexts. Different lexicons have focused on capturing sentiment in diverse genres and domains: from General Inquirer (Stone et al. 1962) categorising words along psycholinguistic dimensions, to VADER (Hutto and Gilbert 2014) optimised for social media; and from MASTER (Loughran and McDonald 2011; Bodnaruk, Loughran, and McDonald 2015) tailored for financial text, to DED (Fioroni et al. 2023) capturing discrete emotions in political communications. However, effectively applying these lexicons in rule-based systems faces several challenges:

- *Disparate, fragmented resources requiring labourious integration*. This inaccessibility restricts replicability, and thus hindering application and advancement of techniques relying on them. The resources are scattered across various sources, such as GitHub repositories, appendices of publications, supplementary materials, and author/institutional websites. Furthermore, they are distributed in diverse file formats, necessitating the tedious process of exporting and importing data into a format compatible with the researcher's workflow.
- *Lack of validated, preprocessed, and high-quality lexicons spanning domains*. Numerous lexicons, including those that undergo peer review, frequently encounter challenges such as the presence of duplicates accompanied by conflicting labels. A substantial portion of existing dictionaries (60%) required removal of duplicates, function words (e.g. prepositions), and rows or columns lacking substantive sentiment content.

To address these limitations, this paper introduces **sentibank** – an integrated, open database consolidating 15 original sentiment dictionaries and 43 processed dictionaries from those originals. Spanning 7 genres and 6 domains, sentibank stands out as the most extensive and comprehensive repository of its kind currently accessible.

The remainder of this paper is organised as follows: First, an overview introduces the sentiment lexicon resources integrated into sentibank, with a focus on the original validation methods. Next, details of the preprocessing techniques applied to standardised dictionaries are presented, followed by exploratory analyses of those preprocessed dictionaries. Finally, potential applications and limitations are discussed, with ethical aspects of the resource.

Related Work

The sentibank dataset contains 15 original dictionaries and 43 preprocessed versions. The source dictionaries include, alphabetically: **AFINN** (Nielsen 2011), **Aigents+** (Raheman et al. 2022), **ANEW** (Bradley and Lang 1999), **Dictionary of Affect in Language (DAL)** (Whissell 1989, 2009), **Discrete Emotions Dictionary (DED)** (Fioroni et al. 2023), **General Inquirer** (Stone et al. 1962), **Henry** (Henry 2008), **MASTER** (Loughran and McDonald 2011; Bodnaruk, Loughran, and McDonald 2015), **Norms of Va-**

lence, Arousal and Dominance (NoVAD) (Warriner, Kuperman, and Brysbaert 2013; Warriner and Kuperman 2015), **OpinionLexicon** (Hu and Liu 2004), **SenticNet** (Cambria et al. 2010; Cambria, Havasi, and Hussain 2012; Cambria, Olsher, and Rajagopal 2014; Cambria et al. 2016, 2018, 2020, 2022), **SentiWordNet** (Esuli and Sebastiani 2006; Baccianella, Esuli, and Sebastiani 2010), **SO-CAL** (Taboada et al. 2011), **VADER** (Hutto and Gilbert 2014), and **WordNet-Affect** (Strapparava and Valitutti 2004; Valitutti, Strapparava, and Stock 2004; Strapparava, Valitutti, and Stock 2006).

While surveying the creation methodologies offers insights, our examination emphasises the validation techniques used for the sentiment dictionaries in sentibank. A key motivation for scrutinising validation is the lack of standardisation in current practices, which can undermine resource quality (Fioroni et al. 2023; Venkit et al. 2023). Many existing lexicons have not undergone thorough validation – out of the 15 dictionaries, 46% employed only single validation techniques – presenting challenges for downstream tasks relying on them. Though this paper predominantly serves as a dataset presentation, clarifying these methods establishes a foundation for enhancing the construction and evaluation of dictionaries in future SA research.

Sentiment Dictionary	Validation Methods
AFINN	Benchmark(ext)
Aigents+	Benchmark(ext)
ANEW	Connotative
DAL	Concordance, Connotative, Contextual, Discriminant
DED	Benchmark(ext), Conceptual, Contextual, Discriminant
General Inquirer	Benchmark
Henry	Conceptual, Contextual
MASTER	Conceptual(ext)
NoVAD	Concordance, Connotative, Contextual, Discriminant
OpinionLexicon	Benchmark
SenticNet	Benchmark(ext)
SentiWordNet	Benchmark(ext)
SO-CAL	Benchmark, Benchmark(ext), Contextual
VADER	Benchmark(ext), Connotative
WordNet-Affect	None

Table 1: Sentiment Dictionaries and Validation Methods. "(ext)" denotes that the validation method included comparison to external sources. For instance, "Benchmark(ext)" indicates performance was evaluated by comparing it to other dictionaries or models on a benchmark dataset. If "Benchmark" is mentioned alone, it signifies that no external comparison was conducted.

In the process of collecting sentiment dictionaries, there were 6 primary validation methods commonly applied: *Benchmark*, *Conceptual*, *Concordance*, *Connotative*, *Contextual* and *Discriminant* validations. The most prevalent were benchmarking against gold-standard corpora through cross-dictionary comparisons. The summary of validation methods employed by each dictionary is presented in Table 1. The remainder of this section will provide an overview of these validation approaches to foster a comprehensive understanding of current practices.

Benchmark Validation

Benchmark validation tests a lexicon's real-world effectiveness by evaluating its performance on manually annotated datasets. Standard metrics like accuracy or F1 are calculated against ground truth labels. For example, Taboada et al. (2011) assessed the accuracy of SO-CAL using a dataset that included a wide range of content from blogs, news articles, and social media corpora. The validation often involves comparing the performance to other dictionaries. For example, Hutto and Gilbert (2014) showed that VADER outperformed 7 other popular dictionaries on multiple benchmarks.

Benchmarking reveals predictive validity on representative tasks with diverse real texts. However, available labelled data may not fully generalise across genres and domains. Complementary validation on niche corpora is advised where suitable benchmarks exist. Still, benchmark tests provide intrinsic assessments using realistic samples of the phenomena lexicons aim to encode.

Conceptual Validation

Conceptual validation involves linking lexicon score patterns to theoretical expectations derived from prior research, providing theory-driven validation. For instance, Fioroni et al. (2023) investigated whether temporal emotion trends align with hypotheses regarding campaign dynamics found in political psychology literature. Often, the lexicons are compared with other dictionaries for further validation: Loughran and McDonald (2011) confirmed MASTER's significant negative correlation with market reactions during 10-K filings, demonstrating robustness compared to the H4N Dictionary.

The conceptual alignment serves as a complement to quantitative validations, ensuring that lexicons effectively capture real-world phenomena as described in conceptual models. However, it is essential to recognise that assumed theoretical knowledge may have limitations, and literature-based protocols should not unilaterally override quantitative evidence.

Concordance Validation

Concordance validation compares a lexicon's annotations against established peer dictionaries to evaluate intrinsic reliability. Correlational analysis between the lexicon and trusted gold-standard dictionaries is common. For example, Warriner, Kuperman, and Brysbaert (2013) investigated correlations of the 'Valence', 'Arousal', and 'Dominance' dimensions with six other existing dictionaries. Statistically significant correlation with widely accepted lexicons indicates validity and standardisation.

Concordance checks offer standardised gauges of quality by assessing alignment with respected peer-designed resources. However, similarly constructed dictionaries may share inherent limitations that inflate correlations. Overreliance on correlational metrics risks perpetuating systemic flaws. Still, judicious cross-dictionary analysis provides pragmatic intrinsic validation.

Connotative (Crowdsourced) Validation

Crowdsourcing sentiment ratings from diverse native speakers helps validate how accurately a lexicon captures cultural connotations. By aggregating scores from multiple raters, idiosyncratic biases tend to average out, approximating population-level associations. This "wisdom-of-the-crowd" provides a reasonable gold standard benchmark to evaluate alignment with collective sociocultural meanings.

For example, Hutto and Gilbert (2014) crowdsourced raters to score words on a 9-point polarity scale, using the mean ratings as valid measures. Further, the authors discarded the entire rating from a rater, if a rater was more than one std away from the mean of the validated sentiment rating distribution.

Of course, biases can arise from non-representative raters or assessment instructions. But on balance, aggregated native speaker ratings offer a practical validation approach for sentiment dictionaries, gauging how well they align with mainstream affective associations. Tapping into collective sociocultural knowledge helps determine if a lexicon accurately captures emotional meanings.

Contextual Validation

Examining lexicon usage in real-world texts from the target domain assesses ecological validity. Contextual validation reveals suitability for practical applications.

For instance, Taboada et al. (2011) highlighted that SO-CAL lexicons were present in 54% of the news headlines dataset, indicating their high term frequency and appropriate usage in corpora. However, niche expressions may not appear, limiting the utility of usage metrics alone. Contextual analysis should combine both qualitative and quantitative examination.

Discriminant Validation

Discriminant validation assesses the distinctness of different sentiment dimensions within a lexicon. As an illustration, Fioroni et al. (2023) calculated correlation between discrete emotion categories to evaluate how well they capture unique affective constructs.

Low correlation implies distinctiveness – the affective dimensions diverge as intended rather than exhibiting redundancy. High discrimination suggests the lexicon encodes distinct affective phenomena rather than conflating into broader polarities like positive/negative.

However, some dimensions may naturally co-vary to an extent in real-world data. Strictly uncorrelated variables could indicate overly contrived constructions not reflecting genuine emotional patterns. Some middle ground must be struck when interpreting discrimination metrics.

Data Collection

sentibank aims to consolidate high-quality sentiment dictionaries that have been widely adopted across SA research. The inclusion criteria focused on dictionaries that have been studied and applied beyond just the SA domain, indicating broad scholarly impact. As shown in Table 2, citation counts were surveyed for prominent sentiment dictionaries using Google Scholar (as of December 2023). For resources with multiple versions or expansions over time, the citation count aggregates papers associated with such dictionary.

While citation count provided an important benchmark, compiling an extensive repository spanning diverse genres and domains was also a key objective. Prior research has noted dependencies in sentiment expression across textual genres like news and social media, as well as topical domains like finance and politics (Remus 2015). For instance, Pennebaker et al. (2015, p. 12) commented on style variations in text across genres. Domain-specific lexicons like MASTER in finance (Loughran and McDonald 2011) also suggest the need for tuning in lexicon sources. By consolidating both widely-adopted general sentiment resources alongside domain-optimised dictionaries, sentibank aimed to enable more robust SA across genres and domains.

Dictionaries were downloaded from various sources, as authors upload using different platforms. Sources included GitHub (Aigents+, SentiWordNet, SO-CAL, VADER), directly from papers and their attached electronic supplementary materials (AFINN, ANEW, DED, Henry, NoVAD), or author/institutional websites (DAL, General Inquirer, MASTER, OpinionLexicon, SenticNet, WordNet-Affect).

Preprocessing

The core contribution of sentibank goes beyond simply consolidating sentiment dictionaries – it applies systematic preprocessing to standardise and quality-check the dictionaries, enabling rapid utilisation. To the best of current knowledge, no other repository offers this combination of curated quality and standardisation[1]. By integrating systematically processed and standardised dictionary resources, this enables researchers to directly apply the unified sentiment data, rather than expending effort formatting and quality checking diverse lexicons.

While no substantial modifications were needed for some lexicons like AFINN, Aigents+, and DED, resources including General Inquirer, Henry, MASTER, OpinionLexicon, SenticNet, SO-CAL, and VADER underwent minor cleaning and formatting to ensure consistency[2]. For example, the SO-CAL dictionary was refined for reusability by excluding 177 algorithm-dependent words and adjusting 269 duplicates. And VADER was refined by averaging 13 duplicates, and adjusting two lexicons with contradictory polarity ratings.

However, more extensive preprocessing was required for ANEW, DAL, NoVAD, SentiWordNet and WordNet-Affect. The key idea behind these preprocessing steps is to mitigate

[1] In the Python ecosystem, there are no libraries integrating diverse sentiment dictionaries. While certain R packages such as quanteda, SentimentAnalysis, and tidytext do offer sentiment dictionaries, they present them without undergoing substantial modification or validation.

[2] Details can be found in doc.socius.org/sentibank

Sentiment Dictionary	Genre	Domain	Licence	Citations
AFINN	Social Media	General	ODbL v1.0	1,752
Aigents+	Social Media	Cryptocurrency	MIT	7
ANEW	Social Media	General	Publicly Available	3,865
DAL	Vernacular	General	Publicly Available	1,043
DED	News	Political Science	Publicly Available	N/A
General Inquirer	General	Psychology	Free for Academic Research	341
Henry	Corporate Communication	Finance	Publicly Available	1,083
MASTER	Regulatory Filings	Finance	Free for Academic Research	5,610
NoVAD	Vernacular	General	Publicly Available	2,061
OpinionLexicon	Product Reviews	Consumer Products	Free for Academic Research	10,308
SenticNet	General	General	MIT	2,806
SentiWordNet	General	General	CC BY-SA 4.0	7,896
SO-CAL	General	General	CC BY-SA 4.0	4,005
VADER	Social Media	General	MIT	5,573
WordNet-Affect	General	Psychology	CC BY 3.0	2,511

Table 2: Overview of Available Sentiment Dictionaries. The term "Publicly Available" indicates datasets sourced from papers or their supplementary materials. "Free for Academic Research" implies instances where original authors explicitly state permissions in their licences. Attribution, providing appropriate credit to associated papers, is a common requirement across most licences.

ambiguities and inconsistencies arising from fuzzy or vector representations by harmonising the representations into well-defined, exclusive schemes[2]. The remainder of this section explains the in-depth modifications made to these dictionaries prior to inclusion in sentibank.

ANEW

The original ANEW dictionary provided normative 'Pleasure', 'Arousal', and 'Dominance' ratings in range of [1,9]. Out of the three emotional dimensions, 'Pleasure' is most directly linked to sentiment polarity. 'Arousal' and 'Dominance' can be more ambiguous - high 'Arousal' may be positive or negative depending on context. Thus, two processed versions were created:

1. **ANEW_v1999_simple**, focusing solely on the pleasure dimension as an indicator of sentiment valence.
2. **ANEW_v1999_weighted**, incorporating all dimensions using a weighted sum.

ANEW_v1999_simple scales mean pleasure ratings (originally ranging from [1,9]) to sentiment scores within the range of [-4,4] using min-max scaling.

ANEW_v1999_weighted incorporates all three dimensions into a single sentiment score using a weighted sum. Though Bradley and Lang (1999) suggested a link between 'Pleasure' and 'Arousal', correlation analysis found these dimensions to be uncorrelated. Instead, 'Pleasure' and 'Dominance' showed a strong positive correlation. Given these relationships, pleasure was assigned a weight of 0.7, 'Dominance' 0.2, and 'Arousal' 0.1 in the weighted sum. The much higher weight for pleasure reflects its stronger direct association with sentiment valence. The non-zero weights for 'Dominance' and 'Arousal' incorporate those dimensions while still emphasising pleasure as the primary driver. These initial weights provide a reasonable starting point, but can be further tuned based on insights into the domain or application. The weighted sum values are scaled using min-max scaling to range from [-4,4].

DAL

DAL rated lexicons on 'Pleasantness' , 'Activation' and 'Imagery' dimensions using a 3-point scale. Since the dictionary was not limited to emotional words, DAL contained many function words such as prepositions (e.g. "when") and pronouns (e.g. "it"). These words had significantly lower imagery scores, indicating they were more abstract (Whissell 2009, p. 513). Words with an imagery score of 1.0 were unlikely to convey emotional meaning, so the 885 words meeting this criteria were removed, leaving 7,858 words in the lexicon.

Two different processed dictionaries have been compiled based on this reduced Dictionary of Affect in Language (DAL):

1. **DAL_v2009_norm** uses a composite sentiment score derived from the 'Pleasantness' and 'Activation' dimensions, drawing on Whissell's (2009) conceptualisation of a two-dimensional affective space.
2. **DAL_v2009_boosted** takes a more experimental approach, exploring an alternative representation of sentiment scores based on 'Pleasantness' and 'Imagery' dimensions. This representation draws on findings from Warriner, Kuperman, and Brysbaert (2013) and Hutto and Gilbert (2014).

Based on prior research on affective spaces (Whissell 2009, p. 510), Whissell highlighted 'Pleasantness' and 'Activation' as the two primary dimensions. In other words, these estimates can be depicted as vector representations, with vector length indicating strength (see Whissell 2009, p. 519). In line with this concept, **DAL_v2009_norm** represents an overall sentiment score using a vector norm that incorporates both 'Pleasantness' and 'Activation' dimensions. The vector norm values are standardised, with scores ranging [-4, 4].

DAL_v2009_boosted combines insights from two additional papers in an attempt to enrich the sentiment encoding. Warriner, Kuperman, and Brysbaert (2013, pp. 1197-1199) reported correlations between 'Valence', 'Arousal', and 'Dominance' with semantic variables like 'Imageability'. Their 'Valence' and 'Arousal' align with DAL's 'Pleasantness' and 'Activation'. This suggested a possibility for representing sentiment scores based on 'Pleasantness' and 'Imagery'.

However, the association between 'Imageability' and 'Valence' was only positive starting from a rating of 5 and higher. Following Hutto and Gilbert (2014), we selectively "boost" the sentiment by adding or subtracting 0.293 with different weights to the 'Pleasantness' score. For words with (scaled) 'Imagery' ≥5 and <5.65, 0.264 (=0.9x0.293) is added/subtracted. For 'Imagery' ≥5.65 and <6.3, 0.278 (=0.95x0.293) is added/subtracted. For 'Imagery' ≥6.3, 0.293 is added/subtracted. This results in scores from [-4.293, 4.293], rescaled to [-4, 4]. However, note that this representation must be used with discretion if 'Imagery' levels are particularly relevant for analysis aims.

NoVAD

Two different processed dictionaries have been compiled for NoVAD:

1. **NoVAD_v2013_norm** represents sentiment as a vector norm encapsulating 'Valence' and 'Arousal'. This approach aligns with Warriner and Kuperman's (2015, p. 16) assertion that an accurate portrayal of sentiment 'requires a bidimensional perspective'.
2. **NoVAD_v2013_boosted** takes a more experimental approach. It seeks to condense sentiment into a single score by adjusting 'Valence' intensity based on 'Arousal' levels.

NoVAD_v2013_norm is created through calculating vector norms of the original 'Valence' and 'Arousal' scores (originally in the range of 1 to 9). This results in min-max scaled scores ranging from -4 to 4.

NoVAD_v2013_boosted refines 'Valence' dimensions based on insights from Warriner and Kuperman (2015). Informed by non-linear relationships identified by Warriner, Kuperman, and Brysbaert (2013), indicating potential threshold effects, adjustments focus on characterising affect using only 'Valence' and 'Arousal'. Warriner, Kuperman, and Brysbaert (2013)'s observation that 'Arousal' modulates 'Valence' suggests considering 'Arousal' as a degree modifier. **NoVAD_v2013_boosted** systematically integrates Hutto and Gilbert's (2014) modifier effect based on Warriner and Kuperman's (2015) chi-squared test analysis.

Specifically, Warriner and Kuperman (2015, pp. 10-11) used a chi-squared test to analyse word types across 100 bins. These bins were formed by the intersection of 10 arousal deciles (A1-A10, with A1 being the lowest arousal percentile and A10 the highest) and 10 valence deciles (V1-V10, with V1 being the most negative percentile, and V10 most positive). This analysis revealed distinct patterns, demonstrating arousal strongly modulates the distribution of word types over valence ratings.

Given these results, we employ a decile-based hierarchy to systematically enhance or dampen sentiment scores. For extreme valence values, we progressively intensify the score based on the arousal decile, amplifying positivity/negativity. Conversely, for neutral valence values, we systematically diminish the score based on the arousal decile, introducing a damping effect to neutrality.

The initial 'Valence' scores underwent a standardisation process, ranging from –4 to 4 through min-max scaling. Subsequently, the polarity of the following word groups was intensified:

- Region [A9:A10, V1] showed chi-square residuals ˜15. This contains highly aroused (top 20%), very negative words (<10%) like "abuse" and "pigheaded". We intensify their negativity by subtracting 0.293. 646 words belong to this lexical space.
- Region [A8, V1] showed residuals ˜10. We intensify negativity of these less aroused (top 30-20%) but still very negative words (>10%) by subtracting 0.278 (=0.95x0.293). 241 words belong to this lexical space.
- Region [A10, V10] showed residuals around ˜10. We intensify the positivity of these highly aroused (>top 10%), very positive words (>top 10%) like "enthusiastic" and "orgasm" by adding 0.278. 239 words belong to this lexical space.
- Region [A7:A10, V2] showed residuals ˜5. We intensify negativity of these moderately aroused (top 40%) but still quite negative words (<20%, ≥10%) by subtracting 0.264 (=0.9x0.293). 797 words belong to this lexical space.
- Region [A7:A9, V10] showed residuals ˜5. We intensify the positivity of these moderately aroused (top 40-10%) yet very positive words (>10%) by adding 0.264. 506 words belong to this lexical space.

The neutrality of the words were also dampened:

- Region [A1:A4, V4:V7] showed residuals ˜5. These are the large lexical spaces, with calmer and relatively neutral words like "foam" and "northern". Indeed, comparing these words with the words in the region [A10, V4:V7], which are those highly aroused but relatively neutral words like "emotional" and "premonition", these are more truly neutral.
- We dampen valence of these words toward neutrality, adding 0.264 for negative words and subtracting 0.264 for positive words, except those already between –0.264 and 0.264. 1,256 words belong to this lexical space.

More negatives (1,684) than positives (745) were intensified, potentially mitigating the positivity bias often discussed in various psychological research (Warriner and Kuperman 2015, pp. 2-5): studies note negative words are less common, and thus perceived more potent; however, positive words appear more frequently, conveying less information. By intensifying negatives, **NoVAD_v2013_boosted** may better reflect how people naturally react to negativity. The higher proportion of intensified negatives counterbalances the typically high informativeness of rare negatives. In total, 2,429 positives/negatives were intensified, 1,256 neutrals were dampened, and 10,230 used the original scaled valence.

SentiWordNet

One distinguishing feature of SentiWordNet, setting it apart from other sentiment dictionaries, is its recognition that *terms can encompass both positive and negative polarities to varying degrees*. For instance, "idle" had an average negative score of 0.375 and positive score of 0.031. This nuanced approach enables SentiWordNet to capture the multifaceted nature of sentiment, acknowledging that words may convey both positive and negative connotations depending on the context in which they are used.

The optimal approach involves embracing polysemy and algorithmically determining positive and negative scores based on the specific context. To capture intended meanings, one could use the Lesk algorithm for word sense disambiguation to assign WordNet synsets to words in context based. This would fully leverage SentiWordNet3.0 in its full potential. However, the purpose of sentibank is to allow researchers to rapidly utilise the processed dictionary. Thus, two different versions of processed dictionaries were created:

1. **SentiWordNet_v2010_simple**, a dictionary that removed ambiguous terms, keeping only strictly positive and negative terms regardless of context; and
2. **SentiWordNet_v2010_logtransform**, a dictionary that retains ambiguous terms using logarithmic transformation for overall scores

For both dictionaries, duplicates were filtered. Of the original 117,659 synsets, 7,031 terms were duplicates across nouns, verbs, adjectives and adverbs. For example, "last" appeared 7 times across 1 noun, 1 adverb, 1 verb and 4 adjective synsets. The 'Pos' and 'Neg' scores for duplicates were averaged. This resulted in 8,636 positive, 62,594 neutral, 9,353 negative, and 5,972 ambiguous terms. Here, ambiguous terms mean terms with non-zero averaged 'Pos' and 'Neg' scores.

SentiWordNet_v2010_simple created by removing all 62,594 neutral and 5,972 ambiguous terms, leaving 17,989 unique terms. The values are scaled using min-max scaling to range from [-4,4].

SentiWordNet_v2010_logtransform involved a logarithmic transformation, specifically defined as

$$\log(Pos + 1) - \log(Neg + 1)$$

This addressed two issues when simply subtracting the scores. Firstly, the transformation serves to mitigate the impact of extreme values in the sentiment scores. Without the logarithmic adjustment, the influence of exceptionally low or high values might overshadow the overall sentiment calculation. Secondly, the logarithmic transformation is adept at preserving the relative differences between positive and negative scores. This ensures that the proportional relationships between scores are maintained, irrespective of their absolute magnitudes. For instance, in a term with a positive score of 0.7 and a negative score of 0.2, the resulting overall sentiment score is 0.546. Conversely, a term with a positive score of 0.5 and a negative score of 0 yields an overall sentiment score of 0.405, demonstrating the preservation of relative differences in the transformed scores.

The transformation resulted in 10,773 positive, 12,130 negative and 63,652 neutral terms. All neutral terms were removed, leaving 22,903 terms. The values are scaled using min-max scaling to range from [-4,4]. It is important to note that **SentiWordNet_v2010_logtransform** increases the coverage of the sentiment dictionary, at the cost of potentially misleading values.

WordNet-Affect

The WordNet-Affect lexicon required extensive preprocessing to consolidate labels and resolve inconsistencies. In particular, the multi-faceted representation differentiating nouns, verbs, adjectives, and adverbs necessitated careful integration to form a coherent resource. The complex attribute hierarchy also demanded reconciling any conflicting inherited labels. The following preprocessing details the key steps taken to transform this pioneering lexicon into a valuable component of sentibank.

1. **Checking Label Alignment**: The attribute labels in the noun-synsets ('a-synsets.xml') were checked for alignment with the attribute hierarchy ('a-hierarchy.xml'). It was unnecessary to check attribute labels for verb-, adjective- and adverb-synsets, as these synsets were semantically linked with appropriate noun synsets. Three noun synset labels – 'joy-pride', 'levity-gaiety', and 'general-gaiety' – did not exist in the hierarchy. These non-existing labels were substituted with the closest matches: 'joy-pride' became 'satisfaction-pride', 'levity-gaiety' became 'playfulness', and 'general-gaiety' became 'merriment'.
2. **Multi-Label Inheritance**: While each synset was originally labelled with one attribute, multiple inherited attributes can be traced in the sentiment hierarchy. For instance, 'peace' belongs to 'tranquillity', which belongs to 'calmness', which finally belongs to 'positive-emotion'. Thus, synset 'peace' inherits multi-labels ['peace', 'tranquillity', 'calmness', 'positive-emotion']. All synsets with original label 'peace' will inherit these sentiment labels.
3. **Re-labelling Multi-Synset Words**: There are four major attribute categories: 'positive-emotion', 'negative-emotion', 'ambiguous-emotion', and 'neutral-emotion'. Often, a word appears in multiple synsets with different attribute categories. And for these multi-synset words, four cases were considered and re-labelled: (i) Contradictory Emotions; (ii) Common Emotions; (iii) Equivocal Emotions; and (iv) Identical Emotions.
   - *Contradictory Emotions*: Words included in synsets with contradictory sentiment labels were reclassified as having ambiguous emotion. For example, "suspense" appeared in synsets n#05583536 and n#05592642, labelled as 'positive-suspense' and 'negative-suspense' respectively. Therefore, "suspense" was relabeled as 'ambiguous-emotion'.
   - *Common Emotions*: Words included in synsets conveying similar sentiments were traced to their common parent emotion in the attribute hierarchy. For instance, "sorrow" occurred in synsets n#05602279 and

n#05601413, tagged with 'regret-sorrow' and 'lost-sorrow'. Both synsets share the parent emotion 'sorrow' in the hierarchy. And thus, "sorrow" was labelled as ['negative-emotion', 'sadness', 'sorrow'].

- *Equivocal Emotions*: Words included in multiple synsets were labelled as ambiguous or neutral emotion if they appeared in both positive/negative synset and ambiguous/neutral synset. For example, "languor" was found in synset n#05563906 with 'neutral-languor' label and n#05587782 with 'positive-languor'. Since "languor" occurred in both a neutral and a positive synset, it was labelled with a more conservative 'neutral-emotion'. In general, if a word was included in an ambiguous or neutral synset, it was marked with an ambiguous or neutral emotion tag.
- *Identical Emotions*: Words labelled with identical sets of emotions were trivially labelled with such sets.

In summary: (i) 4 noun-, 1 adjective-, and 2 verb-synsets had contradictory sets of labels; (ii) 15 noun-, 48 adjective-, 18 verb-, and 1 adverb-synsets had a common set of labels; (iii) 4 noun-, 1 adjective-, and 5 verb-synsets had an equivocal set of labels; and (iv) 1 noun-, 40 adjective-, 23 verb-, and 8 adverb-synsets had an identical set of labels. After relabelling, there were 539 lexicons from the noun-synsets, 609 from the adjective-synsets, 298 from the verb-synsets and 207 from the adverb-synsets.

4. **Re-Labelling Duplicates between Noun, Adjective, Verb and Adverb**: Duplicates between noun-, adjective-, verb-, and adverb-synsets were considered. First, merging noun and adjective synsets yielded 4 duplicates conveying identical, common, and equivocal emotions. Re-labeling these duplicates produced 1,144 lexicons. Next, merging with verb synsets identified 54 duplicates, mostly identical or common emotions, which were relabeled to give 1,384 lexicons. Finally, merging with adverb synsets found just identical emotions. The final **WordNet-Affect_v2006** dictionary resulted in 1,588 unique lexicons after relabeling all duplicates to resolve contradictory labels and consolidate identical or common emotions.

Exploratory Analysis

An essential inquiry involves comparing the positive-negative distributions of dictionaries characterised by categorical labels and by continuous sentiment scores. This analysis was based on the processed versions of the 14 original dictionaries, excluding SenticNet, due to its size. The original 14 dictionaries were categorised into two groups based on their original labelling/scoring frameworks – 8 resources were originally either multi-class or multi-label, and 6 resources were originally continuous.

For the continuous scoring dictionaries, 10 processed variants based on the 6 original dictionaries were analysed. This enabled examination of how different transformations of the original continuous scores impacted resulting distributions. In total, 8 processed label-based dictionaries and 10 processed score-based dictionaries were compared.

Categorical Label Based Dictionaries

To enable a direct comparison among all dictionaries categorising sentiment with labels, labels were simplified into binary categories. AFINN originally has ordinal scores from -5 to +5. Neutral score (0) was excluded, while -5 to -1 were converted to 'Negative' and +1 to +5 to 'Positive'. DED has four discrete emotions, where 'Anger', 'Anxiety' and 'Sadness' were categorised as 'Negative', and 'Optimism' as 'Positive'. For MASTER, 'Uncertainty', 'Litigious', 'Strong Modal', 'Weak Modal', and 'Constraining' were excluded, leaving the 'Positive' and 'Negative' classes. And for WordNet-Affect, 'neutral-emotion' and 'ambiguous-emotion' were excluded, retaining the 'positive-emotion' and 'negative-emotion'.

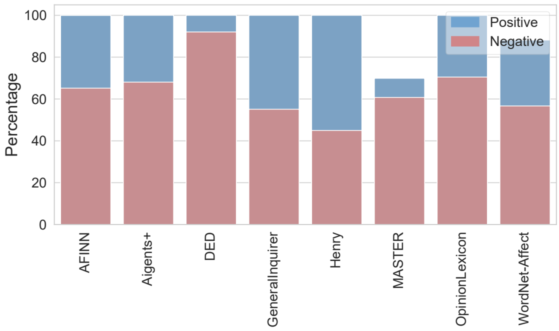


Figure 1: Percentage of Positive and Negative Labels in Categorical Label Based Sentiment Dictionaries

Analysing the percentage of positive and negative labels revealed distinctive sentiment orientations among the dictionaries. As shown in Figure 1, half of the lexicons – AFINN (65.17%), Aigents+ (68.05%), MASTER (60.76%), OpinionLexicon (70.47%) – exhibited a predominantly negative slant. Henry was the sole dictionary with a majority positive sentiment (55.03%).

These class imbalance aligns with how sentiment lexicons were conceptualised and constructed. For instance, AFINN's negative skew can be attributed to its integration of obscene and profane terms from taboo word lists (Nielsen 2011). Meanwhile, Henry (2008, p. 8) noted that companies accentuate positives in earnings reports, explaining the resource's positivity. Conversely, Loughran and McDonald (2011, p. 18) focused on negativity, hypothesising that firms with a high measure of negative words in 10-K filings would experience negative excess returns around the filing date.

Continuous Score Based Dictionaries

Figures 2 and 3 illustrate the divergent distributional properties of sentiment scores across various lexicons. Notably, the distinct distributional shape of SentiWordNet and SO-CAL stemmed from its original scoring scheme: while the sentiment scores fall within the predefined ranges, the values appear to be more akin to ordinal representations with distinct classes, rather than being continuous. This highlights the importance of checking distributional properties when

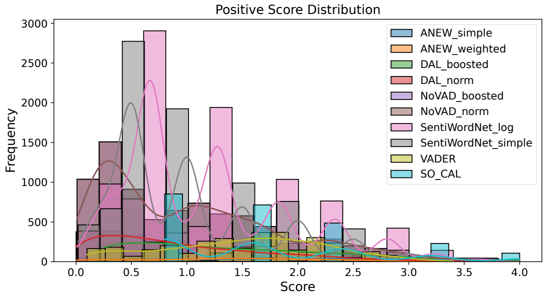


Figure 2: Distribution of Positive Scores Across Continuous Score Based Sentiment Dictionaries

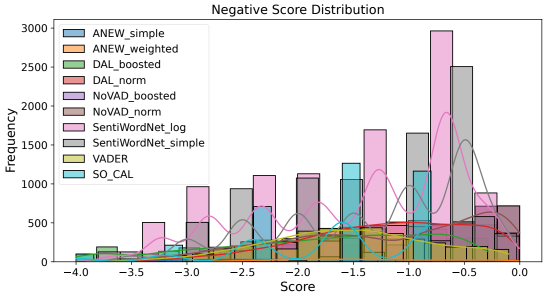


Figure 3: Distribution of Negative Scores Across Continuous Score Based Sentiment Dictionaries

interpreting the results, because the appropriateness of statistical techniques used for analysis may depend on the sentiment scores' distribution, and violated assumptions could lead to invalid conclusions.

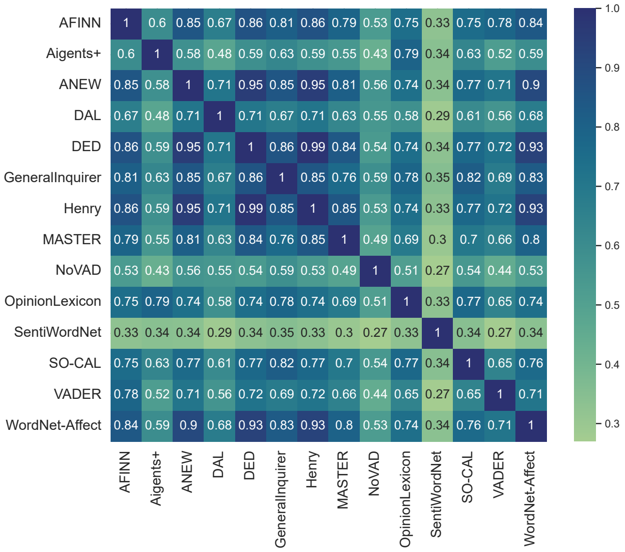
Comparative analysis also revealed distinct variability patterns stemming from different preprocessing methodologies. For instance, weighting affective dimensions in ANEW decreased score variability, potentially due to the dampening effect of incorporating multiple emotional facets. Conversely, DAL_norm and NoVAD_norm displayed positive skewness, compared to their boosted counterparts which exhibited higher variability.

The original scoring schemes by the authors and the preprocessing choices made in sentibank uniquely shape the dictionaries' distributional properties. While no singular optimal sentiment distribution exists, understanding how methodological factors model sentiment expressions provides insights for lexicon selection and interpretation in sentiment analysis tasks.

Jaccard Similarity Analysis

The Jaccard Similarity Matrix (Figure 4) revealed insightful patterns of overlap and dissimilarity between preprocessed sentiment dictionaries when analysed based on their target domains and genres. Comparisons can be categorised into social media, general, and specialised dictionaries.

Dictionaries designed for social media (AFINN-VADER) exhibited moderate similarity (0.783), suggesting considerable but not complete overlap in lexicons optimised for informal contexts like microblogs. General-purpose resources (ANEW, General Inquirer, SentiWordNet, SO-CAL, WordNet-Affect) showed varying yet relatively high similarity, except for SentiWordNet which had lower similarity to all dictionaries. This indicates general-purpose lexicons have substantial overlap but also meaningful differences in composition. Dictionaries for formal genres like news (DED), corporate communications (Henry), and financial filings (MASTER) demonstrated high similarity (0.842-0.987). This implies substantial overlap in terminology for such formal genres. However, the exceptionally high DED-Henry similarity (0.987) may also reflect their small sizes, significantly limiting distinct lexicon options. Still, their shared genre characteristics likely contributes to overlap, despite different finance and political domains.



	AFINN	Aigents+	ANEW	DAL	DED	GeneralInquirer	Henry	MASTER	NoVAD	OpinionLexicon	SentiWordNet	SO-CAL	VADER	WordNet-Affect
AFINN	1	0.6	0.85	0.67	0.86	0.81	0.86	0.79	0.53	0.75	0.33	0.75	0.78	0.84
Aigents+	0.6	1	0.58	0.48	0.59	0.63	0.59	0.55	0.43	0.79	0.34	0.63	0.52	0.59
ANEW	0.85	0.58	1	0.71	0.95	0.85	0.95	0.81	0.56	0.74	0.34	0.77	0.71	0.9
DAL	0.67	0.48	0.71	1	0.71	0.67	0.71	0.63	0.55	0.58	0.29	0.61	0.56	0.68
DED	0.86	0.59	0.95	0.71	1	0.86	0.99	0.84	0.54	0.74	0.34	0.77	0.72	0.93
GeneralInquirer	0.81	0.63	0.85	0.67	0.86	1	0.85	0.76	0.59	0.78	0.35	0.82	0.69	0.83
Henry	0.86	0.59	0.95	0.71	0.99	0.85	1	0.85	0.53	0.74	0.33	0.77	0.72	0.93
MASTER	0.79	0.55	0.81	0.63	0.84	0.76	0.85	1	0.49	0.69	0.3	0.7	0.66	0.8
NoVAD	0.53	0.43	0.56	0.55	0.54	0.59	0.53	0.49	1	0.51	0.27	0.54	0.44	0.53
OpinionLexicon	0.75	0.79	0.74	0.58	0.74	0.78	0.74	0.69	0.51	1	0.33	0.77	0.65	0.74
SentiWordNet	0.33	0.34	0.34	0.29	0.34	0.35	0.33	0.3	0.27	0.33	1	0.34	0.27	0.34
SO-CAL	0.75	0.63	0.77	0.61	0.77	0.82	0.77	0.7	0.54	0.77	0.34	1	0.65	0.76
VADER	0.78	0.52	0.71	0.56	0.72	0.69	0.72	0.66	0.44	0.65	0.27	0.65	1	0.71
WordNet-Affect	0.84	0.59	0.9	0.68	0.93	0.83	0.93	0.8	0.53	0.74	0.34	0.76	0.71	1

Figure 4: Jaccard Similarity Matrix. For SentiWordNet, SentiWordNet_v2010_logtransform was used to represent SentiWordNet as it covers more lexicons than SentiWordNet_v2010_simple.

Applications

1. **Validating New Sentiment Lexicons**: Useful for conveniently cross-validating new dictionary resources without collecting additional labelled data. Particularly valuable when comparing 10+ lexicons, as in Reagan et al. (2016, analysing 24 dictionaries) and Cambria et al. (2022, comparing 20 lexicons across 10 datasets).
2. **Baseline Sentiment Analysis**: Lexicons offer a reproducible, readily-comparable foundation for text-level polarity classification. As Fioroni et al. (2023) discussed, dictionary approaches enable straightforward sentiment scoring, establishing a baseline for contextual and domain-specific refinements. While simplistic in nature, these dictionaries offer a reproducible baseline for capturing essential affect concepts in language, acknowl-

edging the finite nature of affect concepts noted by Cambria et al. (2016).

3. **Interpretability and Ethics**: While deep learning advances SA accuracy beyond simpler lexicons, opaque models undermine trustworthiness in high-stakes decisions (Rudin 2019): for example, deep learning SA models typically reach 70-95% accuracy, surpassing simpler lexicon models in the 55-85% range (Al-Qablan et al. 2023); however, when sentiment analysis is applied in socially impactful domains, transparent and interpretable decision-making is critical. Access to the diverse set of dictionaries and their transparent scoring approaches can potentially assist researchers in comprehending the decisions provided by deep learning driven sentiment analysis models.

Limitations

1. **Overlooking Multidimensional Nature of Emotion**: Representing sentiment along a single dimension risks oversimplifying its nuanced, multidimensional nature expressed in language. Some dictionaries provide rich multidimensional encodings, but condensing into unidimensional scores may overlook key affective signals even when mathematically aggregated. As Warriner and Kuperman (2015, p. 13) noted, focusing on any one affect dimension fails to capture the full complexity of emotional states conveyed through language.
2. **Inherent Subjectivity in Emotional Language Comprehension**: The interpretation of sentiment and emotion in text is inherently subjective and context-dependent. Some texts may subtly convey emotion without overt affective words, while others may contain charged terms but not elicit strong feelings in readers. This nuance poses challenges for dictionary methods focused on explicit polarity terms.
3. **Spelling Variations**: sentibank relied on the original spellings provided within each sentiment lexicon. This was due to dictionaries containing informal language may intentionally include non-standard variants like abbreviations or slang (e.g. ANEW). However, VADER contains usage examples like "aug-00" that do not appear in supplementary dictionaries of abbreviations and slang terms. More rigorous standardisation of these informal spelling variations could better unite lexicons and improve future benchmarking.
4. **Language Scope**: sentibank currently focuses solely on English language dictionaries. Many sentiment resources exist for other languages like German (Waltinger 2010; Remus, Quasthoff, and Heyer 2010), Spanish (Ríos and Gravano 2013; Moreno-Sandoval et al. 2017), French (Abdaoui et al. 2017) and Chinese (Du et al. 2022). Expanding multilingual coverage could increase applicability across global contexts.

Ethical Statement

A recent study by Venkit et al. (2023) highlighted potential sociodemographic biases in SA that could negatively impact real-world applications. Most sentiment lexicons do not consider cultural or contextual differences in how sentiment is expressed. Typically, no checks are done to identify biases that may arise in downstream SA tasks. Analysing such biases is also difficult since sentiment dictionary creators usually only publish final sentiment scores from their sample data, rarely releasing underlying raw data with demographics. Even when demographics are included, the analysis is limited - for example, NoVAD only looked at gender, age and education differences. This prevented sentibank from comprehensively evaluating and mitigating potentially biassed scores/labels, especially for lexicons compiled from social media.

However, unlike opaque deep learning models, sentiment scores are relatively transparent. As an open access resource, sentibank aims to enable community participation that can incrementally improve representation and reduce biases over time. However, each application of these lexicon resources entails an ethical responsibility for mindful usage. Researchers and practitioners building models utilising sentibank are highly advised to carefully consider issues highlighted in the Ethics Sheet for Sentiment Analysis by Venkit et al. (2023, pp. 13750-13751).

FAIR Principles

The sentibank dataset adheres to FAIR principles to maximise its value for current and future research. The FAIR design facilitates integration into diverse workflows, enabling a breadth of applications beyond the initial motivations. The dataset is hosted in a public GitHub repository (https://github.com/socius-org/sentibank), Python Package Index (https://pypi.org/project/sentibank/), and Zenodo (https://zenodo.org/doi/10.5281/zenodo.10514542) under the CC BY-NC-SA 4.0 licence, allowing sharing and adaptation for non-commercial purposes with attribution. The data is provided in standard file formats – CSV, Pickle (dict), and JSON – that are compatible with numerous data analysis applications. As an ongoing project, the most up-to-date information, including preprocessing details and resource summaries, is available at doc.socius.org/sentibank.

References

Abdaoui, A.; Azé, J.; Bringay, S.; and Poncelet, P. 2017. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3): 833–855.

Al-Qablan, T. A.; Mohd Noor, M. H.; Al-Betar, M. A.; and Khader, A. T. 2023. A survey on sentiment analysis and its applications. *Neural Computing and Applications*, 35(29): 21567–21601.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, 2200–2204.

Bodnaruk, A.; Loughran, T.; and McDonald, B. 2015. Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4): 623–646.

Bradley, M. M.; and Lang, P. J. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report 1, Technical report C-1, the center for research in psychophysiology, University of Florida.

Cambria, E.; Havasi, C.; and Hussain, A. 2012. Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *Twenty-Fifth international FLAIRS conference*.

Cambria, E.; Li, Y.; Xing, F. Z.; Poria, S.; and Kwok, K. 2020. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 105–114.

Cambria, E.; Liu, Q.; Decherchi, S.; Xing, F. Z.; and Kwok, K. 2022. SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3829–3839.

Cambria, E.; Olsher, D.; and Rajagopal, D. 2014. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Cambria, E.; Poria, S.; Bajpai, R.; and Schuller, B. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives.

Cambria, E.; Poria, S.; Hazarika, D.; and Kwok, K. 2018. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Cambria, E.; Speer, R.; Havasi, C.; and Hussain, A. 2010. Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI fall symposium series*.

Du, Z.; Huang, A. G.; Wermers, R.; and Wu, W. 2022. Language and domain specificity: A Chinese financial sentiment dictionary. *Review of Finance*, 26(3): 673–719.

Esuli, A.; and Sebastiani, F. 2006. Determining term subjectivity and term orientation for opinion mining. In *11th Conference of the European chapter of the association for computational linguistics*, 193–200.

Fioroni, S.; Hasell, A.; Soroka, S.; and Weeks, B. 2023. Constructing a Dictionary for the Automated Identification of Discrete Emotions in News Content.

Henry, E. 2008. Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, 45(4): 363–407.

Hu, M.; and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

Jawale, S.; and Sawarkar, S. 2020. Interpretable sentiment analysis based on deep learning: An overview. In *2020 IEEE Pune Section International Conference (PuneCon)*, 65–70. IEEE.

Loughran, T.; and McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, 66(1): 35–65.

Moreno-Sandoval, L. G.; Beltrán-Herrera, P.; Vargas-Cruz, J. A.; Sánchez-Barriga, C.; Pomares-Quimbaya, A.; Alvarado-Valencia, J. A.; and García-Díaz, J. C. 2017. CSL: a combined Spanish Lexicon-resource for polarity classification and sentiment analysis. In *International Conference on Enterprise Information Systems*, volume 2, 288–295. SCITEPRESS.

Nielsen, F. Å. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of LIWC2015. Technical report.

Raheman, A.; Kolonin, A.; Fridkins, I.; Ansari, I.; and Vishwas, M. 2022. Social media sentiment analysis for cryptocurrency market prediction. *arXiv preprint arXiv:2204.10185*.

Reagan, A. J.; Tivnan, B.; Williams, J. R.; Danforth, C. M.; and Dodds, P. S. 2016. Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. arXiv:1512.00531.

Remus, R. 2015. *Genre and Domain Dependencies in Sentiment Analysis*. Ph.D. thesis.

Remus, R.; Quasthoff, U.; and Heyer, G. 2010. SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. In *LREC*.

Ríos, M. D.; and Gravano, A. 2013. Spanish dal: a spanish dictionary of affect in language. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 21–28.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Stone, P. J.; Bales, R. F.; Namenwirth, J. Z.; and Ogilvie, D. M. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4): 484.

Strapparava, C.; and Valitutti, A. 2004. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, 40. Lisbon, Portugal.

Strapparava, C.; Valitutti, A.; and Stock, O. 2006. The Affective Weight of Lexicon. In *LREC*, 423–426.

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

Valitutti, A.; Strapparava, C.; and Stock, O. 2004. Developing affective lexical resources. *PsychNology J.*, 2(1): 61–83.

Venkit, P.; Srinath, M.; Gautam, S.; Venkatraman, S.; Gupta, V.; Passonneau, R. J.; and Wilson, S. 2023. The Sentiment Problem: A Critical Survey towards Deconstructing Sentiment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13743–13763.

Waltinger, U. 2010. GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. In *LREC*, 1638–1642.

Warriner, A. B.; and Kuperman, V. 2015. Affective biases in English are bi-dimensional. *Cognition and Emotion*, 29(7): 1147–1167.

Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45: 1191–1207.

Whissell, C. M. 1989. The dictionary of affect in language. In *The measurement of emotions*, 113–131. Elsevier.

Whissell, C. M. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological reports*, 105(2): 509–521.

Paper Checklist

1. For most authors...

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? N/A

(e) Did you describe the limitations of your work? Yes

(f) Did you discuss any potential negative societal impacts of your work? Yes - briefly in Ethical Statement

(g) Did you discuss any potential misuse of your work? No - however, researchers are strongly advised to follow Ethics Sheet proposed by Venkit et al. (2023) when utilising dataset in Ethical Statement

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes - see Ethical Statement

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? N/A

(b) Have you provided justifications for all theoretical results? N/A

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A

(e) Did you address potential biases or limitations in your theoretical framework? N/A

(f) Have you related your theoretical results to the existing literature in social science? N/A

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? N/A

(b) Did you include complete proofs of all theoretical results? N/A

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? N/A

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? N/A

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? Yes

(c) Did you include any new assets in the supplemental material or as a URL? Yes

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? Yes

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? Yes

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? N/A

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A

(d) Did you discuss how data is stored, shared, and deidentified? N/A