# Umigon-lexicon: rule-based model for interpretable sentiment analysis and factuality categorization

Clément Levallois

# umigon-lexicon: rule-based model for interpretable sentiment analysis and factuality categorization

Corresponding author: Clément Levallois, emlyon business school, Paris, France.

Abstract

We present umigon-lexicon, a set of lexicons in English and associated conditions to evaluate sentiment in a text. These lexicons are curated to capture the subjective sentiment conveyed by the author specifically, as opposed to the identification of the overall sentiment. We provide a comparison with existing lexicons before evaluating the performance on sentiment and factuality classification tasks. These contributions highlight the long-lasting relevance of lexicon-based solutions for algorithmic, inherently interpretable models in sentiment analysis and factuality categorization.

Funding declaration: no funding.

Statement of competing interest: the author declares no competing interest.

# 1. Introduction

"Sentiment analysis" or "opinion mining" encompasses a wide range of domains and techniques (Zucco et al., 2020), to the extent that it has been labeled a "mini-version of the full NLP [natural language processing]" (Liu, 2020) or a big "suitcase research problem that requires tackling many NLP tasks" (Cambria et al., 2017). Modern research in the area dates back to the late 1990s (Bradley & Lang, 1999; Wiebe et al., 1999). Initially, researchers in sentiment analysis outlined a research program focused on identifying positive or negative traces of subjectivity and opinions in texts, distinguishing them from impersonal, factual statements (Pang & Lee, 2008). As the field evolved, sentiment analysis diverged into two subdomains: classifying texts as "subjective" or "objective" (Kasmuri & Basiron, 2017; Riloff et al., 2005; Wilson et al., 2005; Yu & Hatzivassiloglou, 2003), and sentiment analysis *per* se, namely detecting the positive or negative sentiment of a text without regard to whether the statement is objective or subjective.

This paper introduces umigon-lexicon, a model designed to integrate these two tasks by focusing on identifying the sentiment expressed in a text as an expression of the author's subjectivity, rather than the sentiment tied to a factual statement or expressed by a third party mentioned in the text. This approach is driven by the recognition that most applications of sentiment analysis, such as categorizing customer reviews or assessing sentiment on social media, require consideration of both sentiment (is the text positively or negatively charged?) and subjectivity (is the sentiment an expression of the author's opinion, or is it a factual statement?). Neglecting either aspect can lead to misclassifications, particularly in cases where the subject matter inherently carries strong negative or positive connotations, such as discussions about tragic events or celebrations (Mohammad, 2021, p. 21). This represents a revival and significant revision of a proposal initially made in 2013 (Levallois, 2013), thoroughly updated to produce the current model.

In the following section, we outline the design principles of umigon-lexicon, highlighting how they compare and contrast with prevalent lexicon-based approaches to sentiment analysis. Subsequently,

we will detail the model's core components in Section 3, followed by an evaluation of its performance in Section 4. We conclude with a discussion of the model's limitations and future prospects.

# 2. Design principles for the lexicons

## 2.1. Identifying the valence of opinions, not facts

Umigon-lexicon aims to characterize the sentiment conveyed by the author of the text rather than capturing the unqualified overall valence of the text. It is essential to clearly distinguish between the valence of facts evoked in the text and the subjective opinions imprinted by the author (Boldrini et al., 2012; Liu, 2010; Mohammad, 2021; Paltoglou et al., 2010; Pang & Lee, 2004; Poria et al., 2023; Tsytsarau & Palpanas, 2012; Wiebe et al., 2004). Often, as in customer reviews, the factual valence and opinion may align, as exemplified by the statement: "I hate that this product broke on the first day, what a shame." In this instance, the negative fact (the product breaking) parallels the negative opinion ("I hate that," "what a shame"). However, conceptually separating the two layers—facts and opinions—reduces bias and aligns more precisely with the definition of "sentiment": "a feeling or an opinion, especially one based on emotions" ("Sentiment (Noun)," 2023, Oxford Advanced Learner's Dictionary), regardless of the inherent positive or negative connotations associated with the subject of the statement (Mohammad, 2021, p. 21). To illustrate, umigon-lexicon is designed with the view to classify item 1 as positive, and item 2 as neutral, in reason of the subjectivity (or lack of) expressed by the author:

1. “I am so glad the peace has been signed” (the author conveying a positive, subjective feeling about the peace being signed)
1. “The peace has been signed” (the author remaining neutral, not sharing their opinion, about the peace being signed even if this utterance has a prior association with a positive sentiment)

This design choice proves advantageous when conducting sentiment analysis studies in specific domains, and more generally when seeking to establish a clear distinction between factual statements

and opinions However, a review of 25 published sentiment lexicons reveals that only four explicitly maintain this distinction (see **Table 1**).[1]

For inclusion in the lexicons, a term must denotate a positive or a negative sentiment. Terms which merely *connotate* a sentiment without conveying a subjective feeling (eg "error", "problem", "war", "peace") are intentionally excluded. When the meaning of a term is context-dependent, umigon-lexicon is structured to allow each lexicon entry to be augmented with conditions and conditional expressions. These additions enable the evaluation of critical aspects of the sentence or proposition where the term appears. Based on this evaluation, a sentiment associated with the term is identified or not (see **Table 2**).

1. "Chocolate is **fantastic**". [Fantastic] denotates a positive sentiment. The term is added to the lexicon.
2. "**Thieves** can go in **prison**": [thieves, prison] connotate a negative sentiment: they have prior associations with a negative sentiment but do not denote it. These terms are not included in the lexicon.
3. "I **dig** this shirt": this statement includes a term [dig] which, without further context, has no certain prior association with positive sentiments. However, the term happens to convey a positive sentiment in reason of being used in the context of a first-person statement ("I dig"). For this precise assessment to be carried out, the term "dig" is added to the lexicon with a condition that the presence of a first-person term in its vicinity should be checked.
4. "He is an **incredible** douche": the term [incredible] can denotate a positive sentiment in some context of use, however as it is followed here by a term denotating a negative sentiment [douche], the positive denotation should be dropped. The term "incredible" is added to the lexicon with a condition making sure that the context of use is considered.

[1] See (Devitt & Ahmad, 2013) for a less extensive but very thorough comparison of lexicons for sentiment analysis.

| | (1) Terms *denotating* sentiment | (2) Terms *connotating* sentiment | (3) Terms not denotating sentiment but context provides it | (4) Terms denotating sentiment but context disconfirms it |
|---|---|---|---|---|
| Umigon | **Yes** | No | **Yes (with conditions affording an assessment of the context of use)** | **Yes (with conditions affording an assessment of the context of use)** |
| Lexicons making no difference between denotations and connotations | **Yes** | **Yes** | No | **Yes** |

**Table 2: Rules for the exclusion or inclusion of terms in the lexicons. Cases (1) (2) (3) (4) are illustrated in the main text.**

## 1.2 Targeting formal and informal writing style

Umigon-lexicon is specifically designed for analyzing short text units, typically spanning a sentence or a paragraph. Social media posts represent a prevalent category of such concise texts, presenting unique analytical challenges. These texts often feature a casual writing style, marked by imperfect syntax, inventive spelling, abbreviations, slang, non-words (such as interjections, emojis, emoticons, hashtags, links, and mentions), and a variety of punctuation marks. A comparative analysis of the vocabulary listed in WordNet 3.0 against several tweet datasets revealed that approximately 45% of the vocabulary in these datasets consists of words not found in the dictionary (Kiritchenko et al., 2014, p. 730).

The unique characteristics of informal text significantly influence the design of lexicon-based systems for sentiment analysis. Techniques such as leveraging syntactic trees (Socher et al., 2013) or label-propagation (Hamilton et al., 2016; Li et al., 2018) achieve optimal results with formal texts, which

adhere strictly to syntactic rules, spelling, and punctuation. However, these methods lose accuracy when applied to short, informal texts that deviate from the semantic structures these models are trained on.

Incorporating out-of-dictionary words enhances the accuracy of sentiment analysis on social media texts. This can be achieved by creating lexicons from a corpus made of such documents written in an informal manner (e.g., tweets) instead of taking dictionaries as a starting point. This corpus-based approach proceeds by first collecting texts likely to be strongly associated with an opinion: movie reviews scored with a one or five stars rating (Pang et al., 2002), product and service reviews associated with terms such as "excellent" or "poor" (Turney, 2002) or that have been recommended or not (Taboada et al., 2011), news articles tagged by readers as evocative of certain emotions (Staiano & Guerini, 2014), tweets marked with sentiment-laden hashtags (Kiritchenko et al., 2014) or emojis associated to different valences (Go et al., 2009). The corpora collected in these manners are then tokenized into their basic elements (unigrams most often, but sometimes bigrams and non-words are also included) and sorted by their frequencies generally measured with a binary count, raw count, or pointwise mutual information. The result of these operations is a lexicon which typically comprises several thousand entries including all the variations that were present in the texts – dictionary words but also misspelled words, slang, hashtags, punctuation signs and more.[2]

This corpus-based approach to creating lexicons excels at covering out-of-dictionary words, since it harvests any token associated with a positive or negative emotion whatever its spelling or meaning,

[2] Given that these lexicons are often created with the purpose to serve as inputs for the training in a model of supervised learning, they can be considered by their authors as a "feature list": a mere by-product or intermediary step in the larger goal of presenting the new design of a model and as such, not always considered worth publishing nor preserving. A noteworthy exception is the effort led by Saif Mohammad at the National Research Council (NRC) Canada who publishes and hosts several such automatically created lexicons. See http://saifmohammad.com/WebPages/lexicons.html

provided it is frequently encountered. However, this method has a significant drawback: it fails to differentiate between the terms that convey specifically the subjective sentiment of the speaker and those which have a prior association with a negative or positive sentiment: all of them will be included provided they are frequent enough. This is a marked inconvenient to achieve a clear distinction between subjective opinions and facts – as detailed in the preceding sub-section. For instance, the NRC Hashtag Sentiment Lexicon includes a list of 54,129 unigrams ranked by their sentiment score (terms associated with a very positive sentiment getting the highest scores) where top-scoring unigrams include “bicycles” and “#realestate”. Such terms do not inherently express the speaker's sentiment, potentially leading to classification errors.

This limitation of corpus-based lexicons led to the design choice of developing umigon-lexicon as a series of manually annotated lists of terms instead. Besides the lexicons for positive and negative words, supplementary lexicons of a smaller size have been created to list negations, conjunctions, and markers of strength of opinions. Various strategies have been employed to integrate out-of-dictionary words:

1. common spelling variations, abbreviations, slang: included in the lexicons.
2. spelling variations such as repeated vowels or consonants (“nooooo” or “so annnnoyed”): handled through an algorithmic pre-treatment procedure on the text, leveraging regular expressions. These forms are detected and transformed into their canonical spelling (“no” and “so annoyed”), which ensures that they will duly be matched in the lexicons.
3. emojis conveying a positive or negative sentiment (👍): detected with lists manually curated, and categorized based on the definition of thematic groups by the Unicode Consortium.
4. Emoticons and their variations (^_^ but also ^_____^): detected and handled with regular expressions.
5. common hashtags (#marryme): development of specific lexicons for these forms.

6. uncommon hashtags where a term conveying a sentiment is present, even when concatenated with another word (#horriblefeeling): development of algorithmic evaluation procedures.

## 1.3 Capturing the context

The valence of the words in a text will contribute to determine the valence of the entire text, however bag-of-words approaches have proven to yield accuracies capped at around 80% (Socher et al., 2013), with a study finding that lexicons alone – without any provision for the effect of valence shifters or else – reached accuracies slightly above or no better than chance (Hartmann et al., 2019). At a minimum, a model for sentiment analysis must address the frequent cases where the polarity of a term is inversed by a valence shifter (if "great" can convey a positive subjective feeling, "not great" probably conveys a negative subjective feeling). A variety of approaches have been explored to accommodate contextual nuances: supervised learning approaches and algorithmic procedures of various degrees of complexity.

One approach in supervised learning consists in capturing contextual effects by learning the statistical associations between features of textual documents associated with a positive or negative valence. This learning is facilitated by using the lexicon entries and their associated valence as a key feature characterizing the documents in the training set (Mohammad et al., 2013). This approach was used in the winning submissions for the sentiment analysis shared tasks of SemEval-2013 Task 2 (Nakov et al., 2013) and SemEval-2014 Task 9 (Rosenthal et al., 2014). More generally, supervised learning approaches can even eschew lexicons altogether and rely only on patterns of relations between tokens and features of a labelled dataset during the training phase of the classifier: tested on a fresh dataset, every token of the text being tested becomes a relevant signal. This approach excels at capturing sentiment with a high precision but cannot retrace or explain its steps in a procedural and interpretable way.

An alternative approach consists in developing algorithmic procedures to evaluate the nature, position, and inter-relations of lexicon entries in the syntactic structure of the sentence. A set of rules

combined in algorithmic routines can be devised to assess the effect of negations, conjunctions, and other constituents of the composition on the sentences that convey a sentiment (see **Table 3**).

| **Reference** | **Number of heuristics** | **rôle of the heuristics** | **Implementation** |
|---|---|---|---|
| (Ding et al., 2008) | An algorithm which includes 12 conditional statements. | Assessing the opinions associated with the features of product mentioned in a product review | None |
| (Hutto & Gilbert, 2014) | 5 generalizable heuristics based on grammatical and syntactic cues. | Evaluating the intensity of an emotion:<br>- punctation (mainly the exclamation point)<br>- capitalization<br>- degree modifiers (intensifiers)<br>- contrastive conjunctions<br>- negations causing valence shifting | VADER |
| (Moilanen & Pulman, 2007) | Dozens of compositional rules derived from *The Cambridge Grammar of the English Language* (Huddleston & Pullum, 2005), *inter alia*. | Recovering the compositional structure of the sentence, which determines how the valence of individual terms impact together the valence of the entire proposition. | None |
| (Choi & Cardie, 2008) | 7 types of polarity shifters | Appreciate better the valence of a sentence where a polarity shifters are used in a variety of configurations | None |
| (Polanyi & Zaenen, 2006) | 10 types of contextual valence shifters. | Early work offering a research agenda to move beyond bag-of-words models. | None |
| (Taboada et al., 2011) | Three. | Detection of intensifiers, negations and irrealis. | SO-CAL |

**Table 3**. Models leveraging deterministic conditional rules to appraise the contextual valence of a term.

umigon-lexicon follows a similar approach by designing a set of conditions which are then selectively associated to *each* lexicon entry by an annotator. Thus, every lexicon entry has been examined

individually and paired with zero, one or several of conditions. While this granular approach adds to the task of the annotators, it allows to very precisely fine-tune how the context should be considered - for every word listed in the lexicons.

As an example, an often-mentioned difficulty in sentiment analysis is the evaluation of valence shifters (Mohammad, 2021; Mohammad & Turney, 2013). While valence shifters in front of positive terms most often yield a negative sentiment ("made my day" conveys a positive sentiment, "didn't made my day" conveys a negative sentiment), the converse is not true: negative terms preceded by a valence shifter might just convey a neutral sentiment ("distressing" conveys a negative sentiment, "not distressing" is neutral rather than positive). Thanks to a condition assessing the presence of a valence shifter on individual words, umigon-lexicon can precisely assess when and how a valence shifter will impact a given positive or negative term. These term-level rules also handle the impacts of:

- moderators of intensity ("very", "absolutely")
- the presence or absence of specific words in the neighborhood ("hard" conveys a negative feeling except when it is immediately followed by the terms "drive", "work", "disk" or is preceded by "party" or "work")
- the presence of capitalized text, subjectivity markers, other words carrying an opinion in the sentence

For a given term in a lexicon, several conditions can be associated in combination, leading to a nuanced appraisal of the context of use of the term. Difficulties well identified in the literature such as expressions with opposite polarities (eg, "happy accidents") (Kiritchenko & Mohammad, 2018) or the shifting role of reinforcers (eg, "this is super!" vs "this is super bad!") can be disambiguated with this approach. These conditions can relate to each other: for instance, the condition evaluating the presence of a valence shifter in the vicinity of a term relies on the detection of moderators of intensity, so that valence shifters are taken into account even when separated of the term by a moderator of

intensity (in the expression “I don’t think he really did a great job”, “don’t” is evaluated as valence shifter for “great” even if it is in the sixth position before the term “great”).

# 2. Description of the lexicons and their conditions

## 3.1. Lexicons

The development of the lexicons was conducted by the author in a low volume but continuous fashion for the last decade. To identify terms solely associated with subjective feelings in umigon-lexicon, we proceeded with an evaluation of a large variety of terms culled from Twitter by studying their denotations, connotations, and actual contexts of use. In the Fall term of 2016, a class project supervised by the author at emlyon business school contributed new terms.[3] From 2013 to 2017, umigon-lexicon was made accessible as a free web application, enabling users to provide feedback on any misclassifications. Since spring 2021, the lexicons have been reintegrated into a web application designed for sentiment analysis, again allowing users to report any errors in classification[4]. The increased traffic to this application has resulted in a greater volume of feedback, significantly accelerating the pace at which the lexicons and their associated conditions are expanded and refined (see **Table 4**).

| | **number of entries** |
|---|---|
| Positive terms | 468 |
| Negative terms | 965 |
| Intensity terms | 168 |
| Valence shifting negations | 119 |

[3] This class project also spurred the creation of lexicons for texts in the French language. An on-going student project is leading to the creation of lexicons for Spanish. These are not covered in this paper.

[4] https://nocodefunctions.com

| | |
|---|---|
| Valence shifting conjunctions | 102 |

**Table 4: Number of entries per lexicon as of November 2023**

A comparison with publicly available lexicons shows that umigon-lexicon is most similar to the AFINN lexicon, the two sets sharing 17% of their terms. This relatively low percentage suggests that umigon-lexicon has still significant margins of improvement in terms of coverage (see **Table 5** for the similarities between 16 popular lexicons).

## 3.2. Conditions

Thirty-eight Boolean conditions are available to evaluate the context of use of a term (see **Table 6**). Not all the conditions are evaluated each time a term is identified in a text. Instead, for each entry in the lexicons, a manual annotation has established which conditions should be evaluated. The true / false value resulting from the evaluation of each condition is fed to an interpreter of conditional expressions (see **Table 7**).

| Lexicon entry | Boolean condition(s) and their parameters | Conditional expression |
|---|---|---|
| lit | isPrecededBySpecificTerm///it's\|its\|it is | 11:10 |

**Table 7: Example of a lexicon entry associated with one condition and a simple conditional expression**

*Legend: if the term "lit" is matched in a text, the following condition will be evaluated: "is the term 'lit' preceded by one of these terms: [it's], [its], [it is]. The result of the evaluation of this condition is a true / false value. A conditional expression ("11:10" here) specifies how this value should translate into a particular valence, codified as: 10 for neutral, 11 for positive, 12 for negative. The value returned by the condition and the conditional expression are then fed to an interpreter. While the interpreter can evaluate multiple conditions associated to complex conditional expressions, the most common case is the simple "which valence when the condition true, and which valence otherwise". Here the valence will be 11 (positive) if "lit" is indeed preceded by [it's], [its], [it is], 10 (neutral) otherwise.*

Conditions return a Boolean value but also metadata, such as the exact position in the text of the terms involved in the evaluation of the condition, which can be leveraged to provide a detailed, transparent description of the operations leading to the result (see also the conclusion).

### 3.3. An engine for the evaluation of conditional expressions

umigon-lexicon utilizes the MVFLEX Expression Language, designed to evaluate any Boolean logic through a syntax analogous to that used in programming languages (Brock & Various contributors, 2021) (see illustration in **Table 8**).

| **A: Lexicon entry** | **B: Boolean condition(s) and their parameters** | **C: Conditional expression** |
|---|---|---|
| heart | isPrecededBySpecificTerm /// break \| breaking \| broke \| broken +++ isFollowedBySpecificTerm /// break \| breaking \| wrenched \| wrenching | if (A \|\| B) {12} else {10} |

**Table 8: Example of a lexicon entry associated with two conditions and a conditional expression involving a Boolean operator and a conditional statement.**

Legend for Table 8:

***In the Boolean conditions column****: the +++ symbol is the separator between two conditions. The /// symbol is used to append parameters to the condition that precedes. The pipe | symbol separates two parameters. By convention, the condition starting the line is "A", the second is "B", etc.*

***In the conditional expression column****: this expression describes how the two conditions of the preceding column should be jointly evaluated, and for what result. The expression will be evaluated by MVFLEX (Brock & Various contributors, 2021), where "A" will be replaced by the Boolean value returned by the first condition, and "B" will be replaced by the Boolean value returned by the second condition.*

*Here, if A OR B are evaluated as "true", MVFLEX will return "12" (which stands for "negative sentiment"), otherwise "10" (which stands for "neutral sentiment").*

Boolean expressions are instrumental in capturing fine contextual cues in exact manner when the valence of a term is particularly dependent of its context of use. Several conditional expressions can be chained (see Table 8, column B) to characterize a given lexicon entry. How the evaluation of each of these conditions yields a result for the given lexicon entry is determined by the conditional expression (Table 8, Column C). This conditional expression is appraised by the evaluation engine. This guarantees that in principle, a variety of contextual effects can be considered to determine the valence of a term accurately.

# 4. Evaluation of quality and performance

## 4.1. Development of quality control through public sharing

umigon-lexicon does not rely on a procedure involving multiple coders to assess the validity of the lexicons and their associated conditions, at the time of their creation. This departure from annotator-based validation methods is primarily due to the cost and complexity associated with the recruitment and training of annotators for extended periods. As discussed in the preceding sections, the distinction is a fine one between a statement which positive valence is stemming from the subjectivity of the author, compared to a statement where the positive valence stems from prior associations while the author remains subjectively neutral Furthermore, the conditions applicable to each lexicon entry, despite being designed for simplicity and ease of learning, inherently lengthen the annotation process. Given these complexities, assembling and maintaining a dedicated team of trained coders for long-term annotation tasks proves impractical, as evidenced by the scarcity of sustained long-term annotation projects in sentiment analysis.

As a viable alternative, leveraging crowd-sourced user feedback has proven efficient. The initial version of the lexicons was created with limited means and without stringent controls, resulting in an imperfect coverage and a limited accuracy. Subsequently, a free web application and API for sentiment

analysis, utilizing these lexicons, were launched, enabling users to report inaccuracies and misclassifications. This approach yields a steady flow of feedback. Crucially, this feedback encompasses a wide array of domains due to the service's open access, devoid of user restrictions. Such contributions are instrumental in broadening the lexicons' coverage, enhancing their external validity, and promoting domain universality.

### 3.4. Measure of performance

umigon-lexicon is designed to identify the sentiment defined as the expression of a subjective expression by the author of a text, with three classes: "positive", "negative" and "neutral". To our knowledge, the MPQA dataset is the only one which maintains this distinction with rigor (Wiebe et al., 2005; Wilson & Wiebe, 2003). It consists in news articles on debated issues on the international political scene. Supplementary information on the dataset can be found in Appendix 1. The test on this dataset was performed against other models of interest (see supplementary information on models in Appendix 2): [5]

1. TimeLMs: a neural language model trained on 123.86 M tweets from 2019 to 2021 (Loureiro et al., 2022). The model scores the highest against competitive models on the sentiment task of TweetEval (Barbieri et al., 2020).
2. GPT-3.5-turbo (Brown et al., 2020): the model enabling the popular application ChatGPT, which scores the highest on a series of advanced cognitive tasks. Two prompts were tested: one "basic" prompt which requests the label of the sentiment of a text and an "advanced" prompt which makes it explicit that it is the subjective sentiment of the author which should be labelled, and neither factuals nor the sentiment of a person mentioned in the text.
3. Mistral 7B, a 7-billion-parameter language model released in October 2023 by the Stanford NLP Group (Jiang et al., 2023). We used the fine-tuned version made available on HuggingFace.

[5] The testbench is made open source with documentation at this link: https://github.com/seinecle/umibench.

The model was tested with “basic” and “advanced” prompts similar to the prompts used with GPT-3-turbo.

| | gpt-3.5-turbo-advanced-prompt | umigon | OpenHermes-2-Mistral-7B-advanced-prompt | gpt-3.5-turbo-basic-prompt | TimeLMs | OpenHermes-2-Mistral-7B-basic-prompt |
|---|---|---|---|---|---|---|
| overall score | 0,867 | 0,860 | 0,847 | 0,827 | 0,762 | 0,708 |
| rank | 1 | 2 | 3 | 4 | 5 | 6 |

**Table 9: Accuracy scores for the test of umigon-lexicon and 3 other models on the MPQA dataset for the task of sentiment analysis. Reported metric is the weighted F1 rounded to 3 decimals**

The evaluation shows that umigon-lexicon ranks second in accuracy, close behind gpt-3.5-turbo when used with a detailed prompt (see **Table 9**). The non-generative AI model included in the set has a noticeably lower accuracy.

umigon-lexicon can also be evaluated on a task of factuality categorization, which has a direct (“upstream”) relation to sentiment analysis (Chaturvedi et al., 2018). Factuality categorization refers to the identification of objective statements, as opposed to subjective ones. As reviewed in the introduction, “sentiment” can usefully be characterized as an expression of the subjectivity of a person. It follows that instances of positive or negative sentiment identified in a text can be categorized as “subjective statements” in a factuality categorization task, and neutral expressions of sentiment can presumably be considered as objective statements.

The test on factuality has been performed with the models for sentiment analysis listed above, except for gpt-3.5-turbo which was too slow to be of use given the volume of entries to test[6]. A model developed specifically for the task of factuality categorization has been added to the bench: “Thesis Titan” (Leistra & Caselli, 2023) which is a multi-lingual model derived from BERT (Devlin et al., 2019)

[6] At the time the tests were performed (November 2023), gpt-3-turbo took about 15 seconds to return an API call.

trained on 2.5 trillion tokens and which scored 3rd at the Clef 2023 workshop on subjectivity detection task.

The datasets to be tested against were purposedly collected from a variety of sources in order to test the models on their domain generalization. Clef 2023 and MPQA are annotated datasets. Alexa, SubQA and X-fact are unlabeled datasets where the label can be inferred by construction. The Kaggle dataset was annotated manually by the author. Supplementary information on each dataset can be found in Appendix 1.

- Alexa: Project by the Amazon Alexa team to evaluate the factual consistency of dialogs (Santhanam et al., 2022). Composed exclusively of objective statements.
- Clef 2023: News articles from the development and training sets of task 2 of the Conference and Labs of the Evaluation Forum (CLEF 2023) (Arampatzis et al., 2023; Galassi et al., 2023).
- Kaggle headlines: a selection of 1,000 entries randomly sampled from a news category dataset collected from 2012 to 202 from HuffPost (Misra, 2022).
- MPQA: the dataset already used for sentiment analysis can also be used to extract labels for objective vs subjective statements
- SubjQA: a selection of customer reviews for the "electronics" product category. Composed exclusively of subjective statements.
- X-fact: a benchmark dataset for multilingual fact checking (Gupta & Srikumar, 2021). Composed exclusively of objective statements.

| | alexa | clef2023 | kaggle-headlines | mpqa | subjqa | xfact |
|---|---|---|---|---|---|---|
| OpenHermes-2-Mistral-7B-advanced-prompt | **0,986** | 0,641 | 0,738 | 0,784 | **0,974** | 0,792 |
| OpenHermes-2-Mistral-7B-basic-prompt | 0,969 | 0,563 | 0,544 | 0,678 | 0,883 | 0,388 |
| Thesis_Titan | 0,964 | **0,821** | 0,857 | **0,877** | 0,789 | 0,960 |
| TimeLMs | 0,872 | 0,610 | 0,719 | 0,706 | 0,948 | 0,671 |
| umigon | 0,962 | 0,602 | **0,944** | 0,781 | 0,957 | **0,977** |

**Table 10: Accuracy scores on the task of factuality categorization. Reported metric is the weighted F1 rounded to 3 decimals**

| | Thesis_Titan | umigon | OpenHermes-2-Mistral-7B-advanced-prompt | TimeLMs | OpenHermes-2-Mistral-7B-basic-prompt |
|---|---|---|---|---|---|
| overall score | 0,921 | 0,920 | 0,797 | 0,701 | 0,503 |
| rank | 1 | 2 | 3 | 4 | 5 |

**Table 11: Overall scores for each model on the factuality task. Scores are the sums of the weighted F1 scores for each dataset, weighted by the number of entries of each dataset.**

The evaluation shows that umigon-lexicon ranks second in accuracy close to Thesis Titan. TimeLMs, which also extends on a BERT model, was trained on a noticeably smaller dataset (135 million tweets) and achieves a lower accuracy (see **Tables 10 and 11**).

## 4. Conclusion: limits and perspectives

Umigon-lexicon demonstrates competitive accuracy across various datasets in sentiment analysis and factuality categorization tasks. Further research efforts need to be conducted in a number of directions.

First, the model is currently incapable of aspect-based sentiment analysis: neither opinion target extraction nor aspect category detection (Do et al., 2019). While a formal research investigation remains to be conducted, it is questionable that an approach based solely on lexicons and their conditional expressions would ever be capable of performing any of these two tasks, which require the model to learn or recover fine information about the syntactical structure of a sentence. Conditional expressions of the sort used by umigon-lexicon may be too crude to capture this information. Integrating a separate model focused on syntactic structure analysis (e.g., Tian et al., 2020) could offer a more effective solution for identifying opinion targets and aspects.

Another research direction involves developing a framework to expand the coverage and performance of lexicons in a principled (and possibly automated) manner. The current approach relies on manual labor supplemented by user feedback (see *supra*). Despite its encouraging results this methodology

remains fragile, hardly reproductible, and is not optimal in terms of coverage. A direction for improvement would consist in automating the discovery of unambiguously subjective statements not currently identified as such by the model, so that they can be submitted to a manual or even an automatic evaluation leading to the addition of new lexicon entries and their associated conditions (Holte, 1993; Letham et al., 2015).

Rule-based models such as umigon-lexicon are inherently interpretable. Provided they do not make use of cognitively taxing methods (such as recursion or overly elaborate Boolean logic) and their number of steps remains modest, a non-expert user is expected to be able to understand the flow of steps chaining an input to an output. This type of model has an important role to play to make decision systems interpretable and auditable (Rudin, 2019). In the domain of sentiment analysis, lexicons augmented with conditions provide the elementary blocks for such an algorithmic procedure. We have made available a web application which generates explanations of the result of the umigon-lexicon by mapping the algorithmic steps followed by the model to their human-readable equivalent.[7]

Lastly, to develop further on the current model and similar lexicon-based models, it is necessary that the research community can easily compare, evaluate, share and build upon the lexicons. For this purpose, umigon-lexicon is made publicly available with a Creative Commons Attribution 4.0 International Public License [8]. Additionally, the lexicons referenced in Table 1 are supplemented by a repository holding a normalized version of these lexicons (with their appropriate licenses) to facilitate their programmatic access[9]. In the case when the license of the lexicons did not allow for their publication, a link is provided to request access.

[7] See the web application cited in footnote 4 above.
[8] Available at https://github.com/seinecle/umigon-family/tree/main/umigon-lexicons.

[9] Available at https://github.com/seinecle/lexicons-for-sentiment-analysis

**References**


Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou, A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., & Ferro, N. (Eds.). (2023). *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, Proceedings* (Vol. 14163). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42448-9

Araque, O., Gatti, L., Staiano, J., & Guerini, M. (2022). DepecheMood++: A Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques. *IEEE Transactions on Affective Computing*, *13*(1), 496–507. https://doi.org/10.1109/TAFFC.2019.2934444

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. LREC 2010, Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification* (arXiv:2010.12421). arXiv. https://doi.org/10.48550/arXiv.2010.12421

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44*(4), 998–1006. https://doi.org/10.3758/s13428-012-0195-z

Boldrini, E., Balahur, A., Martínez-Barco, P., & Montoyo, A. (2012). Using EmotiBlog to annotate and analyse subjectivity in the new textual genres. *Data Mining and Knowledge Discovery*, *25*(3), 603–634. https://doi.org/10.1007/s10618-012-0259-9

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Technical Report C-1 30–1; pp. 25–36). Center for research in psychophysiology, University of Florida.

Bradley, M. M., & Lang, P. J. (2017). *Affective norms for English words (ANEW): Instruction manual and affective ratings* [Technical report C-3]. University of Florida, Gainesville, FL.

Brock, M. & Various contributors. (2021). *MVFLEX Expression Language* (Version mvel2-2.4.14.Final) [Java; Cross-platform]. https://github.com/mvel/mvel

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment Analysis Is a Big Suitcase. *IEEE Intelligent Systems*, *32*(6), 74–80. https://doi.org/10.1109/MIS.2017.4531228

Chaturvedi, I., Cambria, E., Welsch, R. E., & Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, *44*, 65–77. https://doi.org/10.1016/j.inffus.2017.12.006

Choi, Y., & Cardie, C. (2008). Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 793–801. https://aclanthology.org/D08-1083

Deng, L., & Wiebe, J. (2015). MPQA 3.0: An Entity/Event-Level Sentiment Corpus. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1323–1328. https://doi.org/10.3115/v1/N15-1146

Devitt, A., & Ahmad, K. (2013). Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. *Language Resources and Evaluation*, *47*(2), 475–511. https://doi.org/10.1007/s10579-013-9223-6

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 231–240. https://doi.org/10.1145/1341531.1341561

Do, H. H., Prasad, P., Maag, A., & Alsadoon, A. (2019). Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Systems with Applications*, *118*, 272–299. https://doi.org/10.1016/j.eswa.2018.10.003

Galassi, A., Ruggeri, F., Barrón-Cedeño, A., Alam, F., Caselli, T., Kutlu, M., Struß, J. M., Antici, F., Hasanain, M., Köhler, J., Korre, K., Leistra, F., Muti, A., Siegel, M., Türkmen, M. D., Wiegand, M., & Zaghouani, W. (2023). Overview of the CLEF-2023 CheckThat! Lab: Task 2 on Subjectivity Detection. In M. Aliannejadi, G. Faggioli, N. Ferro, & M. Vlachos (Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)* (Vol. 3497, pp. 236–249). CEUR. https://ceur-ws.org/Vol-3497/#paper-020

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision* (CS224N project report). Stanford.

Gupta, A., & Srikumar, V. (2021). X-Fact: A New Benchmark Dataset for Multilingual Fact Checking. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 675–682). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.86

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 595–605. https://doi.org/10.18653/v1/D16-1057

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009

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

Holte, R. C. (1993). Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*, *11*(1), 63–90. https://doi.org/10.1023/A:1022631118932

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177.

Huddleston, R., & Pullum, G. (2005). The Cambridge Grammar of the English Language. *Zeitschrift für Anglistik und Amerikanistik*, *53*(2), 193–194. https://doi.org/10.1515/zaa-2005-0209

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), Article 1.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B* (arXiv:2310.06825). arXiv. https://doi.org/10.48550/arXiv.2310.06825

Jockers, M. (n.d.). Package 'syuzhet'. 2017. *URl: Https://Cran.% 20r-Project.% 20org/Web/Packages/Syuzhet*.

Kasmuri, E., & Basiron, H. (2017). Subjectivity Analysis in Opinion Mining—A Systematic Literature Review. *International Journal of Advances in Soft Computing & Its Applications*, *9*(3), 132–159.

Kiritchenko, S., & Mohammad, S. M. (2018). *Sentiment Composition of Words with Opposing Polarities* (arXiv:1805.04542). arXiv. https://doi.org/10.48550/arXiv.1805.04542

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research*, *50*, 723–762. https://doi.org/10.1613/jair.4272

Leistra, F. A., & Caselli, T. (2023). Thesis Titan at CheckThat! 2023: Language-Specific Fine-tuning of mDeBERTaV3 for Subjectivity Detection. In M. Aliannejadi, G. Faggioli, N. Ferro, & M. Vlachos

(Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)* (Vol. 3497, pp. 351–359). CEUR. https://ceur-ws.org/Vol-3497/#paper-030

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350–1371. https://doi.org/10.1214/15-AOAS848

Levallois, C. (2013). Umigon: Sentiment analysis for tweets based on terms lists and heuristics. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 414–417. https://aclanthology.org/S13-2068

Li, Y., Guo, H., Zhang, Q., Gu, M., & Yang, J. (2018). Imbalanced text sentiment classification using universal and domain-specific knowledge. *Knowledge-Based Systems*, *160*, 1–15. https://doi.org/10.1016/j.knosys.2018.06.019

Liu, B. (2010). Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing* (2nd ed., pp. 627–666). Taylor & Francis. https://www.taylorfrancis.com/chapters/mono/10.1201/9781420085938-36/sentiment-analysis-subjectivity-bing-liu-nitin-indurkhya-fred-damerau

Liu, B. (2020). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108639286

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

Loureiro, D., Barbieri, F., Neves, L., Anke, L. E., & Camacho-Collados, J. (2022). *TimeLMs: Diachronic Language Models from Twitter* (arXiv:2202.03829). arXiv. https://doi.org/10.48550/arXiv.2202.03829

Misra, R. (2022). *News Category Dataset* (arXiv:2209.11429). arXiv. https://doi.org/10.48550/arXiv.2209.11429

Mohammad, S. M. (2017). Word Affect Intensities. *arXiv:1704.08798 [Cs]*. http://arxiv.org/abs/1704.08798

Mohammad, S. M. (2018). Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 174–184. https://doi.org/10.18653/v1/P18-1017

Mohammad, S. M. (2021). Chapter 11 - Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion Measurement (Second Edition)* (pp. 323–379). Woodhead Publishing. https://doi.org/10.1016/B978-0-12-821124-3.00011-9

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 321–327. https://aclanthology.org/S13-2053

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Moilanen, K., & Pulman, S. (2007). Sentiment composition. *International Conference Recent Advances in Natural Language Processing, RANLP*. https://ora.ox.ac.uk/objects/uuid:a03e210a-7734-4059-a2c5-2803c232c10a

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., & Wilson, T. (2013). SemEval-2013 Task 2: Sentiment Analysis in Twitter. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 312–320. https://aclanthology.org/S13-2052

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

Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of Emojis. *PLOS ONE*, *10*(12), e0144296. https://doi.org/10.1371/journal.pone.0144296

Paltoglou, G., Thelwall, M., & Buckley, K. (2010). Online textual communications annotated with grades of emotion strength. *Proceedings of the 3rd International Workshop of Emotion: Corpora for Research on Emotion and Affect*, 25–31.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 271-es. https://doi.org/10.3115/1218955.1218990

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification using Machine Learning Techniques* (arXiv:cs/0205070). arXiv. https://doi.org/10.48550/arXiv.cs/0205070

Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications* (pp. 1–10). Springer Netherlands. https://doi.org/10.1007/1-4020-4102-0_1

Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2023). Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Transactions on Affective Computing*, *14*(1), 108–132. https://doi.org/10.1109/TAFFC.2020.3038167

Riloff, E., Wiebe, J., & Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, 1106–1111.

Rosenthal, S., Ritter, A., Nakov, P., & Stoyanov, V. (2014). SemEval-2014 Task 9: Sentiment Analysis in Twitter. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 73–80. https://doi.org/10.3115/v1/S14-2009

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), Article 5. https://doi.org/10.1038/s42256-019-0048-x

Santhanam, S., Hedayatnia, B., Gella, S., Padmakumar, A., Kim, S., Liu, Y., & Hakkani-Tur, D. (2022). *Rome was built in 1776: A Case Study on Factual Correctness in Knowledge-Grounded Response Generation* (arXiv:2110.05456). arXiv. https://doi.org/10.48550/arXiv.2110.05456

Sentiment (noun). (2023). In *Oxford Advanced Learner's Dictionary*. https://www.oxfordlearnersdictionaries.com/definition/english/sentiment

Sidarenka, U., & Stede, M. (2016). *Generating Sentiment Lexicons for German Twitter* (arXiv:1610.09995). arXiv. https://doi.org/10.48550/arXiv.1610.09995

Siegle, G. J. (1994). *The Balanced Affective Word List Creation Program* [Computer software].

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://aclanthology.org/D13-1170

Staiano, J., & Guerini, M. (2014). Depeche Mood: A Lexicon for Emotion Analysis from Crowd Annotated News. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 427–433. https://doi.org/10.3115/v1/P14-2070

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & 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), 484–498. https://doi.org/10.1002/bs.3830070412

Strapparava, C., & Valitutti, A. (2004, May). WordNet Affect: An Affective Extension of WordNet. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. LREC 2004, Lisbon, Portugal. http://www.lrec-conf.org/proceedings/lrec2004/pdf/369.pdf

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Tian, Y., Song, Y., Xia, F., & Zhang, T. (2020). Improving Constituency Parsing with Span Attention. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1691–1703). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.153

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, *24*(3), 478–514. https://doi.org/10.1007/s10618-011-0238-6

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 417–424. https://doi.org/10.3115/1073083.1073153

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Wiebe, J., Bruce, R., & O'Hara, T. (1999). Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 246–253. https://doi.org/10.3115/1034678.1034721

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning Subjective Language. *Computational Linguistics*, *30*(3), 277–308. https://doi.org/10.1162/0891201041850885

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, *39*(2), 165–210. https://doi.org/10.1007/s10579-005-7880-9

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). OpinionFinder: A system for subjectivity analysis. *Proceedings of*

*HLT/EMNLP on Interactive Demonstrations*, 34–35. https://doi.org/10.3115/1225733.1225751

Wilson, T., & Wiebe, J. (2003). Annotating Opinions in the World Press. *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, 13–22. https://aclanthology.org/W03-2102

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 129–136. https://doi.org/10.3115/1119355.1119372

Zucco, C., Calabrese, B., Agapito, G., Guzzi, P. H., & Cannataro, M. (2020). Sentiment analysis for mining texts and social networks data: Methods and tools. *WIREs Data Mining and Knowledge Discovery*, *10*(1), e1333. https://doi.org/10.1002/widm.1333

# Tables and Figures

## Table 1

A listing of the publicly available lexicons for sentiment analysis. The table provides a comparison on key qualitative aspects of these lexicons. A public repository makes these lexicons available according to their respective licenses: https://github.com/seinecle/lexicons-for-sentiment-analysis

| Dataset | Number of words | Includes ngrams? | Includes out-of-dictionary words? | Includes non-words? | Distinction between subjectivity and facts? | Terms scored by intensity? | License | Commercial use allowed? | Reference |
|---|---|---|---|---|---|---|---|---|---|
| WordNet-Affect | 1417 | yes | no | no | yes | no | Attribution 3.0 Unported (CC BY 3.0) | yes | (Strapparava & Valitutti, 2004) |
| The General Inquirer aka Harvard GI | 3642 | no | no | no | no | no | unclear | unclear | (Stone et al., 1962) |
| MPQA Subjectivity Lexicon aka OpinionFinder | 8222 | no | no | no | yes | no | GNU General Public License | restricted | (Deng & Wiebe, 2015) |

| Dataset | Number of words | Includes ngrams? | Includes out-of-dictionary words? | Includes non-words? | Distinction between subjectivity and facts? | Terms scored by intensity? | License | Commercial use allowed? | Reference |
|---|---|---|---|---|---|---|---|---|---|
| Jockers sentiment | 11710 | no | no | no | no | yes | unclear | unclear | (Jockers, n.d.) |
| AFINN | 2477 | no | no | no | no | yes | Open Database License (ODbL) v1 or a similar copyleft license. | restricted | (Nielsen, 2011) |
| ANEW | 3188 | no | no | no | no | yes | Requestor must be a PhD-holding faculty at a non-profit, degree-granting, academic institution. Sharing is not permitted. | no | (Bradley & Lang, 1999, 2017) |
| VADER | 7520 | no | yes | yes | no | yes | MIT License | yes | (Hutto & Gilbert, 2014) |
| Bing aka Opinion Observer | 6789 | no | no | no | no | no | unclear | unclear | (Hu & Liu, 2004) |
| Sentiment140 and NRC | 740,166 | yes | yes | yes | no | yes | Attribution-NonCommercial 4.0 International (equivalent), full version at: https://saifmohammad.com/WebPages/lexicons.htm | no | (Mohammad et al., 2013) |

| Dataset | Number of words | Includes ngrams? | Includes out-of-dictionary words? | Includes non-words? | Distinction between subjectivity and facts? | Terms scored by intensity? | License | Commercial use allowed? | Reference |
|---|---|---|---|---|---|---|---|---|---|
| Loughran & McDonald | 2,709 | no | no | no | no | no | The dictionary/sentiment lists are free for use in academic research. For commercial licenses, please contact us at loughranmcdonald@gmail.com | no | (Loughran & McDonald, 2011) |
| Henry 2008 | 190 | no | no | no | no | no | unclear | unclear | (Henry, 2008) |
| Emoji sentiment data | 751 | no | no | yes | yes | yes | Attribution 4.0 International (CC BY 4.0) | yes | (Novak et al., 2015) |
| SO-CAL (Semantic Orientation CALculator) | 9,925 | yes | no | no | no | yes | Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License | no | (Taboada et al., 2011) |
| Sentilex | 264 | no | yes | yes | yes | no | MIT License | yes | (Sidarenka & Stede, 2016) |

| Dataset | Number of words | Includes ngrams? | Includes out-of-dictionary words? | Includes non-words? | Distinction between subjectivity and facts? | Terms scored by intensity? | License | Commercial use allowed? | Reference |
|---|---|---|---|---|---|---|---|---|---|
| SentiWordNet | 117,621 | no | no | no | no | yes | Attribution-ShareAlike 4.0 Unported (CC BY-SA 4.0) | yes | (Baccianella et al., 2010) |
| Original Balanced Word List | 277 | no | no | no | no | yes | All included software copyright (c) 1994 by Greg Siegle and the University of Pittsburgh. | no | (Siegle, 1994) |
| Norms of VAD for 13,915 English lemmas | 13,915 | no | no | no | no | yes | unclear | unclear | (Warriner et al., 2013) |
| Checking and bootstrapping lexical norms by means of word similarity indexes | 17,350 | no | no | no | no | yes | unclear | unclear | (Bestgen & Vincze, 2012) |
| Depeche Mood++ | 187,942 | no | no | no | no | yes | This resource can be used for research purposes. Please cite the publications above if you use it. | no | (Araque et al., 2022; Staiano & Guerini, 2014) |

| Dataset | Number of words | Includes ngrams? | Includes out-of-dictionary words? | Includes non-words? | Distinction between subjectivity and facts? | Terms scored by intensity? | License | Commercial use allowed? | Reference |
|---|---|---|---|---|---|---|---|---|---|
| Emolex: NRC Word-Emotion Association Lexicon | 14,154 | yes | yes | no | no | yes | Attribution-NonCommercial 4.0 International (equivalent), full version at: https://saifmohammad.com/WebPages/lexicons.html | no | (Mohammad & Turney, 2013) |
| Word Affect Intensities: NRC Emotion Intensity Lexicon (NRC-EIL) | 5,814 | yes | yes | yes | no | yes | Attribution-NonCommercial 4.0 International (equivalent), full version at: https://saifmohammad.com/WebPages/lexicons.html | no | (Mohammad, 2017) |
| NRC-VAD | 20,007 | yes | yes | yes | no | yes | Attribution-NonCommercial 4.0 International (equivalent), full version at: https://saifmohammad.com/WebPages/lexicons.html | no | (Mohammad, 2018) |

| Dataset | Number of words | Includes ngrams? | Includes out-of-dictionary words? | Includes non-words? | Distinction between subjectivity and facts? | Terms scored by intensity? | License | Commercial use allowed? | Reference |
|---|---|---|---|---|---|---|---|---|---|
| SCL-OPP (Sentiment Composition Lexicon of Opposing Polarity Phrases) | 1,269 | yes | yes | yes | no | yes | Attribution-NonCommercial 4.0 International (equivalent), full version at: https://saifmohammad.com/WebPages/lexicons.html | no | (Kiritchenko & Mohammad, 2018) |
| SemEval-2015 English Twitter Sentiment Lexicon | 1,515 | yes | yes | yes | no | yes | Attribution-NonCommercial 4.0 International (equivalent), full version at: https://saifmohammad.com/WebPages/lexicons.html | no | (Rosenthal et al., 2014) |
| NRC Hashtag Sentiment Lexicon | 370,660 | yes | yes | yes | no | yes | Attribution-NonCommercial 4.0 International (equivalent), full version at: https://saifmohammad.com/WebPages/lexicons.html | | (Kiritchenko et al., 2014) |
| umigon-lexicon | 1,434 | yes | yes | yes | yes | no | Attribution 4.0 International (CC BY 4.0) | yes | this publication |

# Table 5

A comparison of major public lexicons. The percentage represents the cardinality of the two sets, discounted by the size of the largest of the two sets.

| comparing lexicons | Compass DeRose | Norms of valence | umigon-lexicon | Jockers sentiment | Harvard GI | Henry 2008 | Loughran & McDonald | SO-CAL | AFINN | ANEW | bootstrapping lexical norms | wordnet-affect | Original Balanced Word List | MPQA | Bing aka Opinion Observer | VADER | SentiWordNet |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Compass DeRose | | 3% | 10% | 5% | 8% | 2% | 4% | 6% | 23% | 3% | 15% | 9% | 7% | 6% | 8% | 2% | 2% |
| Norms of valence | 3% | | 4% | 37% | 20% | 1% | 6% | 23% | 9% | 59% | 4% | 2% | 16% | 21% | 14% | 18% | 18% |
| umigon-lexicon | 10% | 4% | | 7% | 10% | 2% | 6% | 9% | 17% | 3% | 12% | 3% | 10% | 9% | 8% | 2% | 2% |
| Jockers sentiment | 5% | 37% | 7% | | 25% | 1% | 13% | 34% | 20% | 31% | 7% | 1% | 36% | 56% | 30% | 18% | 18% |
| Harvard GI | 8% | 20% | 10% | 25% | | 1% | 16% | 41% | 24% | 18% | 11% | 3% | 38% | 36% | 17% | 8% | 8% |
| Henry 2008 | 2% | 1% | 2% | 1% | 1% | | 4% | 1% | 3% | 1% | 1% | 3% | 1% | 1% | 1% | 0% | 0% |
| Loughran & McDonald | 4% | 6% | 6% | 13% | 16% | 4% | | 13% | 22% | 6% | 4% | 2% | 12% | 15% | 12% | 3% | 3% |
| SO-CAL | 6% | 23% | 9% | 34% | 41% | 1% | 13% | | 17% | 21% | 10% | 2% | 43% | 50% | 24% | 11% | 11% |
| AFINN | 23% | 9% | 17% | 20% | 24% | 3% | 22% | 17% | | 7% | 14% | 6% | 18% | 19% | 32% | 4% | 4% |
| ANEW | 3% | 59% | 3% | 31% | 18% | 1% | 6% | 21% | 7% | | 4% | 1% | 14% | 19% | 12% | 22% | 5% |
| bootstrapping lexical norms | 15% | 4% | 12% | 7% | 11% | 1% | 4% | 10% | 14% | 4% | | 4% | 14% | 10% | 8% | 3% | 22% |
| wordnet-affect | 9% | 2% | 3% | 1% | 3% | 3% | 2% | 2% | 6% | 1% | 4% | | 2% | 2% | 2% | 0% | 3% |
| Original Balanced Word List | 7% | 16% | 10% | 36% | 38% | 1% | 12% | 43% | 18% | 14% | 14% | 2% | | 59% | 21% | 9% | 0% |
| MPQA | 6% | 21% | 9% | 56% | 36% | 1% | 15% | 50% | 19% | 19% | 10% | 2% | 59% | | 30% | 12% | 9% |
| Bing aka Opinion Observer | 8% | 14% | 8% | 30% | 17% | 1% | 12% | 24% | 32% | 12% | 8% | 2% | 21% | 30% | | 8% | 12% |
| VADER | 2% | 18% | 2% | 18% | 8% | 0% | 3% | 11% | 4% | 22% | 3% | 0% | 9% | 12% | 8% | | 8% |
| SentiWordNet | 2% | 18% | 2% | 18% | 8% | 0% | 3% | 11% | 4% | 5% | 22% | 3% | 0% | 9% | 12% | 8% | |
| **Maximum value for cardinality** | **23%** | **59%** | **17%** | **56%** | **41%** | **4%** | **22%** | **50%** | **32%** | **59%** | **22%** | **9%** | **59%** | **59%** | **32%** | **22%** | **22%** |

# Table 6

List of the Boolean conditions which can be associated with lexicon entries

| **umigon-lexicon (2023)** |
|---|
| is all caps |
| is first letter capitalized |
| is first term of text |
| is followed by a negative opinion |
| is followed by a positive opinion |
| is followed by specific term |
| is hashtag starting with affective term |
| is hashtag |
| is hashtag negative sentiment |
| is hashtag positive sentiment |
| is hashtag start |
| is immediately followed by a negation |
| is immediately followed by a negative opinion |
| is immediately followed by a positive opinion |
| is immediately followed by an opinion |
| is immediately followed by a negative prior association |
| is immediately followed by a positive prior association |
| is immediately followed by a time indication |
| is immediately followed by specific term |
| is immediately preceded by a negation |
| is immediately preceded by an opinion |
| is immediately preceded by positive opinion |
| is immediately preceded by a negative prior association |
| is immediately preceded by a positive prior association |
| is immediately preceded by specific term |
| is immediately preceded by a strong term |
| is immediately preceded by a subjective term |
| is in a sentence with one of these specific terms |
| is in hashtag |
| is in sentence ending in exclamation |
| is last ngram of sentence-like fragment |
| is negation in all caps |
| is preceded by opinion |
| is preceded by positive opinion |
| is preceded by specific term |
| is preceded by strong word |
| is preceded by subjective term |
| is question mark at end of text |

**Appendix 1: supplementary information on the datasets used in the test of the accuracy for umigon-lexicon.**

Datasets were collected with the view to:

- diversify domains (news, reviews, headlines)
- rely on datasets that are labelled according to precise standards for subjectivity and sentiment.

In particular, we did not include datasets where the subjective vs factuality dimension could not be ascertained in a principled way. For instance, the classic subjectivity dataset on subjectivity in movie reviews (Pang & Lee, 2004) which includes many subjective statements in the "plot" (objective) category.

## 1. Clef2023 conference

The corpus is annotated for OBJECTIVITY vs SUBJECTIVITY. Corpus in English of subtask-2-english: dev_en.tsv *and* train_en.tsv datasets.

Labels on factuality:

| objective | subjective |
|---|---|
| 638 (59%) | 411 (38%) |

## 2. News Category Dataset

The corpus is annotated for OBJECTIVITY vs SUBJECTIVITY.

1,000 headlines extracted at random. The author then annotated them for factuality vs subjectivity, as manual inspection revealed that a few headlines could be said to be subjective rather than factual (eg, "Reporter Gets Adorable Surprise From Her Boyfriend While Live On TV" -> coded as "subjective" because of the token "adorable")

Labels on factuality:

| objective | subjective |
|---|---|
| 894 (89%) | 106 (11%) |

## 3. MPQA dataset (1.2)

The corpus is annotated for OBJECTIVITY vs SUBJECTIVITY. Subjective statements are annotated for POSITIVE, NEGATIVE OR NEUTRAL sentiment.

- Entries labelled as "subjective" are the ones annotated with the following annotations: GATE_expressive-subjectivity AND nested-source="w"
- Within the "subjective" entries, sentiment was characterized based on the annotation "polarity", which takes the following values: "positive", "negative", "neutral".
- Entries labelled as "objective" are the ones annotated with the following tag: GATE_objective-speech-event. All "objective" entries were marked as "neutral" for sentiment.
- In all cases, an entry was not included in the dataset if it included the following annotation: polarity="uncertain"

Labels on sentiment:

| positive | negative | neutral |
|---|---|---|
| 57 (2%) | 266 (8%) | 2,873 (90%) |

Labels on factuality:

| objective | subjective |
|---|---|
| 2,501 (78%) | 696 (22%) |

Note : the figure for the “subjective” class does not equal the sum of the “positive” and “negative” classes, because annotators of the MPQA dataset have introduced a class for subjective statements which are neutral.

In the task for sentiment analysis, these neutral, subjective statements are collapsed with the objective, neutral statements into a unique “neutral” class which comprises 2,873 statements.

In the task for factuality categorization these neutral, subjective statements are collapsed with the positive and negative (subjective) statements into a unique “subjective” class which comprises 696 statements.

## 4. SubjQA: A Dataset for Subjectivity and Review Comprehension

The corpus is annotated for OBJECTIVITY vs SUBJECTIVITY. In practice, all entries are labelled as SUBJECTIVE.

In the dataset, we selected the "electronics" product category from the "train" set and filtered the data entries to keep only those which were rated for maximum subjectivity (score of 1 out of 5 in the field "ans_subj_score"). This returns 1,373 records out of 2,346.

Labels on factuality:

| objective | subjective |
|---|---|
| 0 (0%) | 1,373 (100%) |

## 5. X-FACT: A New Benchmark Dataset for Multilingual Fact Checking

The corpus is annotated for OBJECTIVITY vs SUBJECTIVITY. In practice, all entries are labelled as OBJECTIVE.

The entries were extracted from the field "claim" from the entries in English in the file "train.all.tsv" in the "x-fact/data/x-fact-including-en/" directory.

Labels on factuality:

| objective | subjective |
|---|---|
| 12,294 (100%) | 0 (0%) |

## 6. Alexa: Factual consistency analysis of dialogs

This is a project by the Alexa research team on identifying factual correctness.

In the file factual_dataset_expert.csv; the field "knowledge" contains factual statements derived from the field "context".

Entries in this field are objective by construction, not subjective.

In practice, the author extracted all the entries of the "knowledge" field, with filters on two additional fields:

- "hallucination" = No
- "verifiable" = y (yes)

This results in 470 entries.

Labels on factuality:

| objective | subjective |
|---|---|
| 470 (100%) | 0 (0%) |

**Appendix 2: supplementary information on the models used to benchmark the accuracy of umigon-lexicon.**

## 1. TimeLMs

A model dedicated to evaluating the sentiment in texts.From the repository on HuggingFace: "This is a roBERTa-base model trained on ~58M tweets and finetuned for sentiment analysis with the TweetEval benchmark. This model is suitable for English".

Model tested on HuggingFace with this inference endpoint:

https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

## 2. Open Hermes 2 Mistral 7b

A large language model with broad capabilities. From the repo on HuggingFace: "OpenHermes 2 Mistral 7B is a state of the art Mistral Fine-tune. OpenHermes was trained on 900,000 entries of primarily GPT-4 generated data, from open datasets across the AI landscape. [More details soon]".

Model tested on HuggingFace with this inference endpoint:

https://huggingface.co/teknium/OpenHermes-2-Mistral-7B

A non negligible number of entries were misclassified due to the model being unable to perform the task: the instructions in the prompt (basic or advanced) were ignored. Instead of returning a label, the response consisted in a text which was an attempt at continuing the text of the entry provided as an input.

### a) Basic prompt

You are the equivalent of a human annotator. You must:

- label the sentiment of the text provided below. The label should be a single word: "positive", "negative" or "neutral".
- extract the unique identifier of the text. The unique identifier is the string of characters that is at the start of the text, up to the # character.

Your response should be exactly the unique identifier of the text, followed by a space, followed by the label. Do not add the text itself or anything else.

The text: <insert the text to analyze>

### b) Advanced prompt

This is a textbook about natural language processing (NLP). Sentiment analysis is a classic task that we detail in this chapter. The annotation for sentiment consists in labelling a text with one of these three labels: "positive", "negative" or "neutral". For example, the text "I am very happy that she could come" will be labelled as "positive". It is important to note that the quality of the labelling

depends on the strict following of these instructions: - the annotator should use a single word for the label of the sentiment, without further comment. The word should be "neutral", "positive" or "negative". - the annotator identifies a sentiment when the text reflects personal feelings, tastes, or opinions. - the annotator should label the sentiment expressed by the author of the text, not the sentiment expressed by a person cited in the text. - the annotator should be careful that a factual, even when it has strong positive or negative prior associations (such as "war" or "happyness"), is not a sentiment.

The following examples will illustrate this lesson:

Example 1:

- The text to label for sentiment: "I love chocolate"
- The label: positive

Example 2:

- The text to label for sentiment: "She says she loves chocolate"
- The label: neutral

Example 3:

- The text to label for sentiment: <insert the text to analyze>
- The label:

### 3. Thesis Titan

A model dedicated to evaluating the factuality of a text. From the repository on HuggingFace: "Fine-tuned mDeBERTa V3 model for subjectivity detection in newspaper sentences. This model was developed as part of the CLEF 2023 CheckThat! Lab Task 2: Subjectivity in News Articles." "The model ranked third in the CheckThat! Lab and obtained a macro F1 of 0.77 and a SUBJ F1 of 0.79."

Model tested on HuggingFace with this inference endpoint:

https://huggingface.co/GroNLP/mdebertav3-subjectivity-english

### 4. GPT3.5 turbo

A large language model with broad capabilities.

It runs on the MPQA dataset (3,198 entries) for less than a dollar, in approximately 12 hours. Less than 5 entries generated an error and were lost.

Model tested on the API endpoint of OpenAI, with default parameters.

#### a) Basic prompt

Role "system" : You are a the equivalent of a human annotator in a data labelling task. The task consists in labelling the sentiment of a text provided by the user. The label should be a single word: "positive", "negative" or "neutral".

Role "User" : "The text to label: \n\n" + <insert the text to analyze>

### b) Advanced prompt

Role "system" : You are a the equivalent of a human annotator in a data labelling task. The task consists in labelling the sentiment of a text provided by the user. When annotating, be especially attentive to these 3 recommendations:

1. you should annotate the sentiment expressed by the author of the text, not the sentiment expressed by a person cited in the text.
2. a sentiment is expressed when the text reflects personal feelings, tastes, or opinions.
3. a factual, even when it has strong positive or negative prior associations (such as "war" or "happyness"), is not a sentiment.

The label should be a single word: "positive", "negative" or "neutral".

Role "User" : "The text to label: \n\n" + <insert the text to analyze>