
ELECTRA: GENERALIZING LEXICONS IN SENTIMENT ANALYSIS USING REVERSE-SENTIMENT TRANSFORMATIONS

Simon, Rahsaan
rasimon.connect@gmail.com

ABSTRACT

A simple new method is proposed and investigated for generalizing lexicons using Valence, Arousal, Dominance theory to instruct the computer to reverse the sentiment of a word in a lexicon, or perform a "Reverse-Sentiment Transformation", should the word's VAD qualities match a particular designation that is based on subject matter. This technique was primarily investigated with a modified version of the VADER lexicon and algorithm, referred to here as "Electra", and was found to generally result in the increase of the model's overall correctness and the balancing of its statistical precision and recall.

1 Introduction

Sentiment analysis is a subject of great interest in many domains of application and research. Sentiment analysis technologies act as a way of bridging consumers to manufacturers in industry, and their improvement serves as a direct avenue for making accurate affective computing technologies in research. Thus, the improvement of these technologies is of both theoretical and practical interest.

Technological development is currently split into two major avenues: improvement via the use of machine learning and improvement outside of this field via algorithmic means. Improvement within machine learning is straightforward, as artificial intelligence methods are now advanced enough to conduct analyses of sentiment-oriented text with the use of proper training. However, improvement outside of this subject is made possible and competitive by the fact that people naturally assign sentiment to texts algorithmically, and replicating this process is much more efficient than the use of artificial intelligence. The exact determination of this process, however, remains a mystery at the intersection of linguistics, philosophy, and computation.

In spite of their potential competitiveness, modern lexicon-based approaches tend to struggle in domains outside of expertise, a challenge that is easily overcome by competing machine learning counterparts. However, the computational complexity of these systems, which are quite tedious to train and expensive to execute, makes a search into alternative methods a vital subject.

Current lexicon-based approaches tend to manipulate the prescribed valences of individual words in a way that depends on the context of the sentence in which the word is used, often termed as a set of *golden rules* which improve the accuracy of models based on an attempt at replicating how humans perceive the sentiments of sentences and paragraphs. Often, though, these rules are ambiguous and, in their execution, apply well to one domain but perform noticeably worse in others.

Modern research then aims to formulate a set of golden rules which are verbose enough to cover a wide-range of subjects, but are simple enough so that the approach remains computationally inexpensive, at least to a point where it is worth the use of lexicon-based approaches in favor of other methods for their efficiency, even if they are slightly less reliable.

Within context, it is possible to view the task of generalization as binary, wherein a word in another context takes a binary shift from positive to negative while maintaining its emotional magnitude (for example, "killer" as a compliment is meant to be just as impactful as "killer" in a formal context, only positive). Additionally, many of these words, which are changed within a context, can be viewed as very emotionally related, by some arbitrary designated components e.g., killer and radical.

Emotional qualities can be split into components via the prominent Valence, Arousal, Dominance (VAD) model of emotion, wherein each combination of VAD values represents its own emotion, with neighboring values representing related emotions. Therefore, words that have a strong emotional connection will have the same or a similar set of VAD values.

With these two facts, it is then possible to create a true *golden rule* which encompasses other golden rules by allowing for the binary switching of words of a certain VAD combination from negative to positive or positive to negative in a specified context.

We introduce a system compatible with existing lexicon models that accomplishes this simply via a brute-force search of all VAD combinations and the accuracy they induce, and as a result, see a noticeable increase in the models' capability for generalization across multiple domains.

The emotions of words in this schema have been accrued in the crowd-sourced NRC VAD Lexicon.

2 Methodology

2.1 Reverse-Sentiment Transformations

The "Reverse-Sentiment Transformation" is the core technique of Electra, wherein a term with a VAD array (as is determined by the term's lexicon entry) that matches another given array, which is referred to as the "transformation array", has its valence value reversed in polarity, such that it effectively acts as its own sentimental antonym.

The conception of this technique was motivated by the inherent differences between two terms in two different contexts. Take, for instance, the term "scary": whereas the term is generally negative, in the domain of film critiques, it is generally positive to the same degree that it was negative, especially if the genre of review is horror. In this scenario, scary can be identified and separated from the rest of the lexicon via its VAD array, in a manner such that similar terms with a similar concept (e.g. thrilling) are also caught and reversed in sentiment. In this way, a model performs better, on average, since it has recognized an important emotional aspect of the topic it is reviewing, which could not have been possible with logic alone.

In fact, one could go so far as to say that this opens up the possibility for a new regime that is entirely different from logical improvements, and potentially allows for new improvements that could allow for simulated pattern recognition that is comparable to artificial intelligence models.

The following is an example of how this technique would be implemented in code via a very simple inference function taking in singular reviews and returning their average polarity (the sum of terms' polarity divided by the number of terms; terms that are neutral are not considered, as is standard throughout the rest of this paper)

2.2 Lexicon

For the purposes of this research, a modified version of the VADER lexicon [1] was used, with its values being the source of the valence values of its various terms. Sources for terms' arousal and dominance values were taken from

the NRC VAD dataset [2] [3] and were rounded to one significant figure for grouping.

Terms within the VADER lexicon that did not have an equivalent in the NRC dataset were given placeholder arousal and dominance values such that they would be unapplicable to any specific transformation array and unaffected by any resultant transformation.

2.3 Fine-Tuning

The specific VAD array that would be used to determine whether a transformation is relevant and significant to the context was determined by testing all possible combinations of VAD transformation arrays of a certain order within a basic algorithm, with the inclusion of an extra value (1.1) that represents compatibility with any value of that type. For this project, values of 0-1 with increments of order 0.1 were used, such that possible values were 0, 0.1, 0.2, ... 1.1.

This method can be seen as being anomalous to fine-tuning with machine learning models, with the only exception being that the best found transformation array (of a specified order) should be generally applicable to all such cases wherein text from the array's prescribed context is being analyzed. This alleviates the need to further search for transformation arrays for a context with algorithmic improvements and model developments, making transformation arrays independent of model architecture, but heavily dependent on the lexicon.

Below is a table of the transformation arrays, found by fine-tuning, that preempt the results that appear in this paper.

Context	Best Transformation Array (V, A, D)
Movies	(0.7, 0.3, 1.1)
Products	(0.7, 0.3, 1.1)
Social Media	(0.7, 0.6, 0.6)
Finance	(0.3, 0.5, 0.2)

2.4 Classifications

Electra performs binary classifications. Final valence values that are above 0.5 are interpreted as positive, while values below 0.5 were interpreted as negative, and values equal to 0.5 were discarded.

All datasets used were datasets that either had binary classifications or were edited to only include binary classifications.

3 Results

Results for electra in comparison to vader differ across subject matter, although all subjects displayed an increase in overall correctness (the amount of labels the model

Generalizing Lexicons

correctly predicted) when using Electra and reverse-sentiment transformations.

Additionally, and quite generally, algorithm variants that used reverse-sentiment transformations often balanced

precision and recall values compared to those that did not, proving that the model provided more reliable classifications for the particular datasets to which they were fine-tuned.

Movie Reviews				
Model (Version)	Overall	Precision	Recall	f1 Score
basic (w/o transformations)	16759 / 25000	0.646	0.871	0.742
basic (w/ transformations)	17710 / 25000	0.724	0.764	0.744
vader (w/o transformations)	17497 / 25000	0.651	0.862	0.742
vader (w/ transformations)	18150 / 25000	0.712	0.761	0.735

Product Reviews				
Model (Version)	Overall	Precision	Recall	f1 Score
basic (w/o transformations)	2633 / 4000	0.633	0.943	0.758
basic (w/ transformations)	2755 / 4000	0.693	0.836	0.757
vader (w/o transformations)	2706 / 4000	0.628	0.934	0.751
vader (w/ transformations)	2810 / 4000	0.676	0.829	0.745

Financial News				
Model (Version)	Overall	Precision	Recall	f1 Score
basic (w/o transformations)	677 / 1356	0.808	0.894	0.849
basic (w/ transformations)	708 / 1356	0.785	0.932	0.852
vader (w/o transformations)	706 / 1356	0.779	0.888	0.830
vader (w/ transformations)	719 / 1356	0.774	0.927	0.844

Microblogs				
Model (Version)	Overall	Precision	Recall	f1 Score
basic (w/o transformations)	12364 / 25000	0.659	0.877	0.753
basic (w/ transformations)	12518 / 25000	0.671	0.872	0.758
vader (w/o transformations)	13054 / 25000	0.660	0.858	0.746
vader (w/ transformations)	13202 / 25000	0.670	0.852	0.750

4 Discussion

The results are a promising demonstration of the increase in accuracy transformations bring to existing algorithms, as well as an indicator of the potential transformations have to add to existing lexicon techniques, wherein they can work collaboratively with adjustments in rule-based

methods to make them more accurate for a context rather than going against them.

Additionally, a rather surprising result is that the effects of transformations tend to balance recall and precision

Generalizing Lexicons

values, showing more stable predictions compared to the models without.

The most influential results for the field of affective computing research and sentiment analysis will most certainly be the highest performing transformation arrays, which may be used for the respective context so long as the lexicon remains unchanged. In addition to this usefulness, they might also be a decent indicator of how certain words with emotional qualities change depending on the context, on average. For example, for the movies context, based on the fact that the best transformation array consists of the values 0.7, 0.3, and 1.1, it is possible to definitively state that all words which have a valence of 0.7, arousal of 0.3, and dominance of 1.1, within the NRC lexicon should be modified to best fit the context of movies (based on the used dataset). With more data and research, it may be possible to take this hypothesis a step further and deem that all such words in general, regardless of lexicon and dataset used for training, should have their valence modified to appropriately represent human emotion for any

given subject, thus providing a definitive method for generalizing sentiment analysis algorithms and lexicons and demonstrating the utility of VAD theory.

References

- [1] C. Hutto and Eric Gilbert. 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):216–225, May 2014.
- [2] Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018.
- [3] Saif M. Mohammad. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms, 2025.