Confirming the Non-compositionality of Idioms for Sentiment Analysis

Alyssa Hwang
Computer Science Department
Columbia University
a.hwang@columbia.edu

Christopher Hidey
Computer Science Department
Columbia University
chidey@cs.columbia.edu

Abstract

An idiom is defined as a non-compositional multiword expression, one whose meaning cannot be deduced from the definitions of the component words. This definition does not explicitly define the compositionality of an idiom's sentiment; this paper aims to determine whether the sentiment of the component words of an idiom is related to the sentiment of that idiom. We use the Dictionary of Affect in Language augmented by WordNet to give each idiom in the Sentiment Lexicon of IDiomatic Expressions (SLIDE) a component-wise sentiment score and compare it to the phrase-level sentiment label crowdsourced by the creators of SLIDE. We find that there is no discernible relation between these two measures of idiom sentiment. This supports the hypothesis that idioms are not compositional for sentiment along with semantics and motivates further work in handling idioms for sentiment analysis.

1 Introduction

The processing of multiword expressions (MWEs) is an underrecognized subfield of natural language processing research. A multiword expression is defined as a phrase that can be decomposed into multiple lexemes and shows lexical, syntactic, semantic, pragmatic, or statistic idiosyncrasy (Baldwin and Kim, 2010), where a lexeme is a linguistic unit that constitutes the basic block of a language (Ramisch, 2015). MWEs are prevalent in modern text and increasing in frequency as modern language develops—Jackendoff (1997) estimates that the number of MWEs in a speaker's lexicon is roughly equivalent to the number of single words, and 44% of entries in WordNet 3.0 are multiword (Miller, 1995), a 3% increase from WordNet 1.7 (Sag et al., 2002). Ignoring MWEs when analyzing natural speech can result in models that cannot handle variation or fail to generalize, and relying on complicated preprocessing or ad hoc methods of handling MWEs creates systems that are difficult to maintain or extend (Sag et al., 2002).

Idioms, a subset of MWEs, are particularly challenging to analyze because they are non-compositional: the meaning of the entire idiom cannot be deduced from the definitions of each individual word in it (Jochim et al., 2018). Treating idioms like "it's raining cats and dogs" with a words-with-spaces approach can diminish the accuracy of a model that treats each word as the smallest unit of a sentence; the example idiom simply means that it is raining heavily and is unrelated to animals. Along with meaning, past work has already shown that ignoring idioms in sentiment analysis tasks will lower the accuracy of a sentiment classifier (Williams et al., 2015), but the non-compositionality of idiom sentiment is not included in the currently acknowledged definition of an idiom and should not be immediately assumed without further research.

The goal of this paper is to confirm or deny the non-compositionality of idiom sentiment. Some idioms, like "a blessing in disguise," "so far so good," "in the red," and "add insult to injury," show potential compositionality of sentiment based on the positive sentiments of "blessing" and "good" and negative sentiments of "red," "insult," and "injury." Other examples, like "break a leg," "speak of the devil," and "let the cat out of the bag," would imply the wrong sentiment based on the negative sentiment in "break" and "devil" and lack of strong polar sentiment in any of the words "let," "the," "cat," "out," "of," and "bag." Based on the definition of an idiom, that the collective meaning of component words does not predict the meaning of the entire phrase, we hypothesize that the sentiment of an idiom is non-compositional. We test this hypothesis by comparing two scores for each idiom in the Senti-

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 125–129
Florence, Italy, August 2, 2019.

ment Lexicon of IDiomatic Expressions (SLIDE): a DAL sentiment score based on each word in the idiom and a SLIDE positive percent index given by the lexicon.

2 Related Work

Williams et al. (2015) explore how much the inclusion of idioms as features improve traditional sentiment classification and provide a set of 580 idioms annotated with sentiment polarity and a corpus of sentences containing idioms in context. Each sentence was labeled with an emotion and the authors compared models that predicted the gold standard by including and excluding separate treatment of idioms. When comparing the results, they noted significant improvement in F-score for all three sentiment classes: positive, negative, and other. The results of Williams et al.'s work demonstrate the need to include additional methods for handling idioms in sentiment analysis.

Ramisch and Villavicencio (2018) define the linguistic characteristics of MWEs and discuss how to incorporate MWEs into language technology. Savary et al. (2017) produce a multilingual 5-million-word annotated corpus of verbal MWEs (such as "to *break* one's heart") and annotation guidelines for eighteen languages. Seretan (2008) provides a syntax-based methodological framework for automatically identifying idiomatic collocations in text corpora. Many neural models of sentiment, like the one used by Socher et al. (2013), assume that sentiment is compositional. Zhu et al. (2015) incorporate both compositional and non-compositional sentiment by using an automatic labeling method for the non-compositionality of n-grams while we focus on annotated idioms.

Jochim et al. (2018) present SLIDE, the Sentiment Lexicon of IDiomatic Expressions. SLIDE is a collection of 5,000 idiomatic expressions, a great expansion from Williams et al.'s set of 580 idioms. Jochim et al. used CrowdFlower to have at least ten annotators label each idiom as positive, negative, neutral, or inappropriate. The lexicon includes the distribution of annotations and a sentiment label that represents the label that received the majority of votes. In the case of a tie between positive/negative and neutral, the idiom is labeled positive/negative; in the case of a tie between positive and negative, the idiom is labeled neutral. The SLIDE polarity annotations were critical for the endeavors of this paper.

To compute sentiment scores for idioms based on each component word, we relied on the technique developed by Agarwal et al. (2009) to detect phrase-level polarity. They derived lexical scores for pleasantness, activation, and imagery from the Dictionary of Affect in Language (M. Whissel, 1989) augmented by WordNet (Miller, 1995), used a finite state machine to handle local negations, and boosted scores to capture the strength of words that may have otherwise received similar pleasantness scores—consider the difference between "fairly good advice" and "excellent advice," for example. We implemented their method of computing sentiment scores to compare to phrase labels provided by SLIDE.

3 Methods

3.1 SLIDE Positive Percent Index and Sentiment Label

We used the Sentiment Lexicon of IDiomatic Expressions (SLIDE) (Jochim et al., 2018) to give each idiom a positive percent index and sentiment label. The sentiment labels were given by the lexicon as a majority vote of at least ten crowdsourced annotations per idiom, and only idioms that are labeled positive (946), negative (1,108), or neutral (2,945) were used in this study, for a total of 4,999 idioms. The full dataset was used for analysis. The positive percent index was calculated by subtracting the percentage of negative votes from the percentage of positive votes. This system of quantitatively evaluating sentiment emphasizes the positive score of an idiom without distinguishing neutral and negative sentiment. In this study, we focus on positive sentiment; alternatives include calculating negative or neutral percent indices or subtracting just the negative percentage of votes to capture the nuances of sentiment strength.

3.2 Component-wise Idiom Scoring

We compute component-wise scores by implementing Agarwal et al.'s method of measuring phrase-level polarity (Agarwal et al., 2009). These scores represent the compositional sentiment of an idiom. We begin by tokenizing the idiom (Honnibal and Montani, 2017) and assigning each word a pleasantness score from the Dictionary of Affect in Language (DAL) (M. Whissel, 1989); if the word is not present in the DAL, we use the pleasantness score for a synonym or the negated pleasantness score for an antonym from WordNet (Miller, 1995). We consider each word sense from WordNet in order, which is based on the frequency of use, and use the first sense that had a DAL entry. The scores are Z-normalized according to the mean and standard deviation of each sentiment class given in the manual for the DAL and boosted by multiplying by the number of standard deviations they lie from the mean.

We then handle local negations with a finite state machine of two states: RETAIN and INVERT. The scores remain the same when the finite state machine is in the RETAIN state and are negated when in the INVERT state. Each idiom starts in the RETAIN state and switches to the INVERT state when a negation, like "not," "no," and "never," is encountered. The finite state machine returns to the RETAIN state if it encounters the word "but" or a comparative degree adjective, like "better" or "worse," to account for phrases like "no better than evil." The idiom's component-wise score is the sum of the scores for each component word normalized by the length of the idiom.

4 Results and Discussion

We have computed the Spearman correlation between the predicted and gold labels and p-values for each sentiment class, with the null hypothesis that two sets of data are not correlated. The Spearman correlation of each sentiment class is close to 0, which implies no correlation, and we fail to reject the null hypothesis for idioms labeled neutral and negative. Even though $p \leq 0.05$ for idioms labeled positive, the near-zero Spearman correlation of -0.144 still indicates no correlation between predicted and gold labels. These values further support our claim that idioms are non-compositional for sentiment.

	Spearman corr.	p-value
Positive	-0.144	9.35×10^{-6}
Neutral	0.012	0.503
Negative	0.007	0.813

Table 1: Spearman correlation scores and p-values

When plotted against the crowdsourced sentiment distribution from SLIDE, the component-wise sentiment scores show no obvious pattern (see Figure 1). In total, 19% of idioms were labeled positive, 22% labeled negative, and 59% labeled neutral.

The SLIDE positive percent indices range from -1.0, which means that no annotators labeled the idiom positive, to 1.0, which means that all annotators labeled it positive. Figure 1 shows clear separation between idioms labeled positive (○) and idioms labeled negative (□) but does not distinguish between negative and neutral (×), as expected. It does, however, show the lack of obvious correlation between the crowdsourced positive percent index (horizontal axis) and computed DAL positive index (vertical axis).

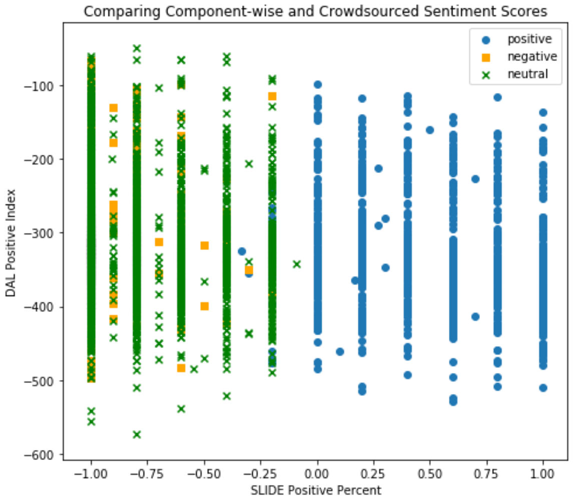

Figure 1: Component-wise sentiment score vs. SLIDE positive percent index with sentiment labels

Table 1 below contains a few examples of idioms with varying scores computed from the DAL. It shows how idioms with the same label can have widely varying scores from SLIDE and the DAL and provides empirical evidence for the non-compositionality of idiom sentiment.

Idiom	Label	PPI	DAL
Two thumbs up	Positive	0.8	-377
Get one's feet wet	Positive	0.2	-197
Fifth wheel	Negative	-1.0	-293
Third degree	Negative	-1.0	-309
Word-for-word	Neutral	-1.0	-254
Let it be	Neutral	-1.0	-288

Table 2: Examples with sentiment labels, positive percent index (PPI), and DAL positive index (DAL)

Figures 2, 3, and 4 show the range in component-wise and SLIDE sentiment scores for each polarity class: positive, negative, and neutral. If idioms were compositional for sentiment, we would expect SLIDE positive percent and DAL

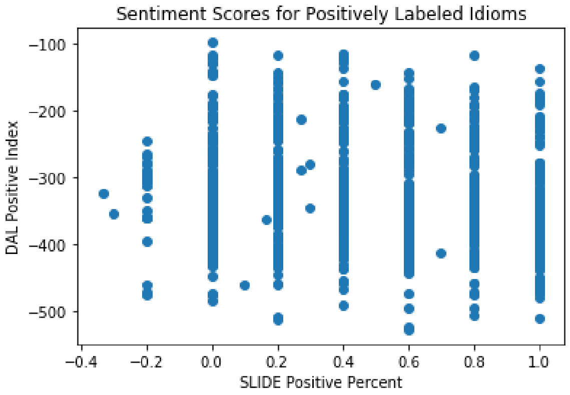

Figure 2: Component-wise and SLIDE sentiment scores for idioms labeled positive. $n = 946$, DAL mean: -328.68, DAL std: 78.44

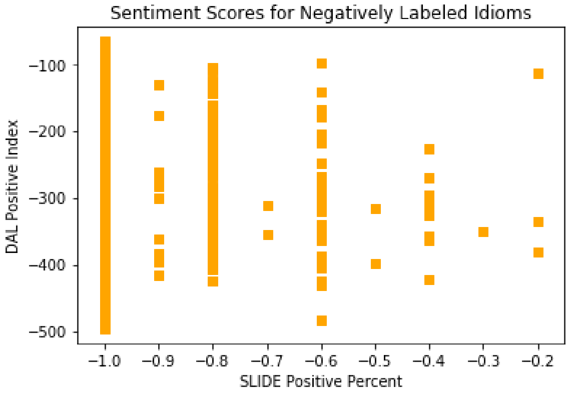

Figure 3: Component-wise and SLIDE sentiment scores for idioms labeled negative. $n = 1108$, DAL mean: -274.90, DAL std: 66.16

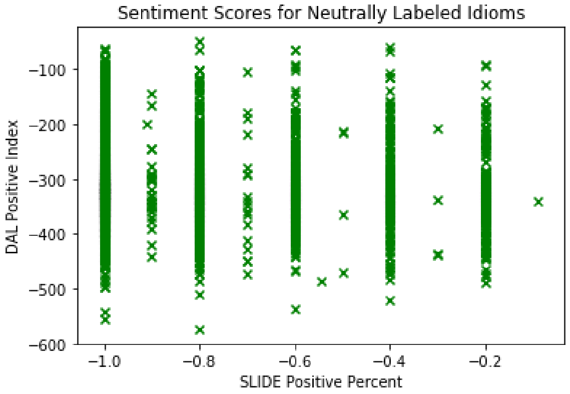

Figure 4: Component-wise and SLIDE sentiment scores for idioms labeled neutral $n = 2945$, DAL mean: -57.63, DAL std: 17.72

positive index to be directly related, but we can see from Figure 1 that idioms with the highest SLIDE positive percent rating do not strictly correspond to a higher DAL positive index. In fact, there seems to be no relationship between SLIDE positive percent and DAL positive index at all. In Figure 1, we can see no distinct pattern between the two measurements of phrase sentiment.

Furthermore, even though the SLIDE positive percent index poorly distinguishes between idioms with majority negative and neutral votes, we would expect to see consistently lower DAL positive indices for idioms labeled negative than idioms labeled neutral. Negatively labeled idioms do have a noticeably lower mean DAL positive index but a much larger standard deviation than neutral idioms. Surprisingly, positively labeled idioms have an even lower mean DAL positive index than negatively labeled idioms, with a comparable standard deviation. It is interesting that negatively and positively labeled idioms (idioms that express some emotion) both display much lower mean values and much greater standard deviations of DAL positive index scores while neutral (unemotional) idioms tend to vary less. This may indicate that emotional idioms contain emotional words, but the sentiment of the words does not necessarily correlate to the sentiment of the entire phrase.

5 Conclusion and Future Work

Our analysis shows that there is no consistent correlation between component-wise sentiment scores and crowdsourced phrase-level labels, which supports the hypothesis that idioms are non-compositional for sentiment as well as meaning. The non-compositionality of sentiment was not explicitly defined or immediately obvious for idioms, and the lack of relationship between component words and phrase-level sentiment motivates further research in handling idioms in context. Multiword expressions in general are very common and increasing in frequency in modern language, and we have demonstrated that treating MWEs as words-with-spaces rather than separate, complete entities can lead to inconsistent results in sentiment labeling.

Possible future work in the sentiment analysis of MWEs include learning domain-specific sentiment without manual annotation, like predicting a negative sentiment for the phrase “high blood pressure” in the context of a poor health condition. Work must also be done in recognizing new MWEs as language evolves, as well as associating new meanings to already existing words and phrases. This is particularly important for processing Internet slang, which evolves and generates new vocabulary very quickly through social media. For example, the saying "yeet haw," a combination of the words "yeet" and "yeehaw," which are both casual expressions of excitement, has risen in occurrence. Manually annotating common idioms, as the creators of SLIDE had CrowdFlower workers do, is a tedious, time-consuming, and never-ending task as long as language keeps changing. Learning to recognize and associate proper sentiment scores to MWEs is an important step in improving overall sentiment classification.

6 Acknowledgments

We would like to thank Kathy McKeown from Columbia University for her generous support and the reviewers for their thoughtful feedback.

References

Apoorv Agarwal, Fadi Biadsy, and Kathleen McKeown. 2009. Contextual Phrase-Level Polarity Analysis Using Lexical Affect Scoring and Syntactic N-Grams. In *EACL*.

Timothy Baldwin and Su Nam Kim. 2010. *Multiword Expressions*, pages 267–292.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *To appear*.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. Linguistic Inquiry Monographs. MIT Press.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. SLIDE - a Sentiment Lexicon of Common Idioms. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Cynthia M. Whissel. 1989. *The Dictionary of Affect in Language*, pages 113–131.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

Carlos Ramisch. 2015. *Definitions and Characteristics*, pages 23–51. Springer International Publishing, Cham.

Carlos Ramisch and Aline Villavicencio. 2018. Computational Treatment of Multiword Expressions. *The Oxford Handbook of Computational Linguistics 2nd edition*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Violeta Seretan. 2008. *Collocation Extraction Based on Syntactic Parsing*. Ph.D. thesis.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasic. 2015. The Role of Idioms in Sentiment Analysis. *Expert Systems with Applications*, 10.

Xiaodan Zhu, Hongyu Guo, and Parinaz Sobhani. 2015. Neural Networks for Integrating Compositional and Non-compositional Sentiment in Sentiment Composition. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 1–9, Denver, Colorado. Association for Computational Linguistics.