

AWESSOME: An Unsupervised Sentiment Intensity Scoring Framework using Neural Word Embeddings

Amal Htait^[0003-4647-9996] and Leif Azzopardi^[0002-6900-0557]

University of Strathclyde, Glasgow, UK
{amal.htait@strath.ac.uk, leif.azzopardi@strath.ac.uk}

Abstract. Sentiment analysis (SA) is the key element for a variety of opinion and attitude mining tasks. While various unsupervised SA tools already exist, a central problem is that they are lexicon-based where the lexicons used are limited, leading to a vocabulary mismatch. In this paper, we present an unsupervised word embedding-based sentiment scoring framework for sentiment intensity scoring (SIS). The framework generalizes and combines past works so that pre-existing lexicons (e.g. VADER, LabMT) and word embeddings (e.g. BERT, RoBERTa) can be used to address this problem, with no require training, and while providing fine grained SIS of words and phrases. The framework is scalable and extensible, so that custom lexicons or word embeddings can be used to core methods, and to even create new corpus specific lexicons without the need for extensive supervised learning and retraining. The Python 3 toolkit is open source, freely available from GitHub ¹ and can be directly installed via *pip install awesome*.

Keywords: Sentiment Intensity · Pre-trained Language Model · Lexicon · BERT · VADER.

1 Introduction

With the increasing usage of social media platforms, there has been great interest from various sectors, such as sociology, psychology, and marketing, to analyse and monitor such streams in order to extract people’s opinions, attitudes and emotions [2, 6]. Consequently, numerous sentiment analysis (SA) techniques and methods have been proposed and developed over the years [3, 6, 11]. Given the daily streams of posts, tweets, blogs and reviews, where people talk about products, places, people, etc, the aim of SA techniques has been to either classify the content as positive, neutral or negative (Sentiment Classification), or to rate the intensity of the sentiment on a scale from strongly positive to strongly negative (Sentiment Intensity Scoring (SIS)). The later being the harder task, which can then be used for classification purposes. The extremely informal nature of online texts varies significantly from formal texts, creating challenges for traditional

¹ <https://github.com/cumulative-revelations/awesome>

unsupervised SA techniques, which rely mainly on direct keyword matching and scoring using a highly curated sentiment lexicon (e.g. VADER [2], LabMT [1], LIWC [9]) combined with a series of crafted rules. While, these dictionary based approaches required no training and play an essential role in the fast and scalable analysis of large volumes of online posts, they are fundamentally limited. This is due to the vocabulary mismatch problem, as the vocabulary of the target text is different from the sentiment lexicons, reducing the effectiveness of methods trying to score the sentiment intensity of the phrase or sentence. This work aims to develop a framework for building SIS methods that: (1) addresses the vocabulary mismatch problem by using neural word embeddings, (2) capitalises on pre-existing lexicons for validated sentiment scores, (3) runs out of the box, without requiring any supervised training, and (4) is configurable, customisable and scalable.

1.1 Related Tools and Lexicons

While there are many approaches that have been proposed which are based on supervised machine learning, our focus is on unsupervised approaches. Largely unsupervised approaches are dictionary based - that is they draw upon a curated lexicon of {word, sentiment score} pairs - which are used in conjunction with hand-crafted rules: **SentiStrength**² [10] is a sentiment strength extraction tool, it classifies text based on a dual 5-point system for positive and negative sentiment. For that purpose, SentiStrength uses a sentiment dictionary and employs a range of well known non-standard spellings in addition to other common textual methods of expressing sentiment. **VADER** [2] is a widely used simple rule-based model for SA. It relies on a sentiment dictionary (7500 records) of gold-standard quality with human-validated valence scores that indicated both the sentiment polarity (positive/negative), and the sentiment intensity on a scale from -4 to +4 (e.g. *good* has a positive valence of 1.9, *great* is 3.1). VADER's sentiment lexicon was compared to seven well-known SA lexicons and it proved its well performance, particularly in the social media domain [2]. Another well known sentiment lexicon, **LabMT** [1] of 10,222 words with their average happiness (or positivity) calculated by combining word frequency distributions and an independently assessed numerical estimates of the *happiness* of over 10,000 words obtained using Amazon's Mechanical Turk³. An embeddings based approach, **ASID**⁴ [4], was more recently proposed to score SIS. In their approach the sentence words were matched against positive and negative seed lists' terms via corpus based word embeddings, and then the average difference between the positive and negative lists were used to produce the final sentiment intensity score. In this work, we extend this approach to draw upon the lexicons previously developed and validated and to combine them with pre-existing neural word embeddings to create a generalized framework for SIS.

² <http://sentistrength.wlv.ac.uk/>

³ <https://www.mturk.com/>

⁴ <https://github.com/amalhtait/ASID>

2 The AWESSOME Framework

The presented framework, *A Word Embedding Sentiment Scorer Of Many Emotions* (AWESSOME), has the purpose of predicting the sentiment intensity of words and sentences. The generalised framework is inspired by our previous work [4], which relies on using sentiment seed terms and word embeddings, where the similarity between the vector representation of two sentences is considered as a reflection of their sentiment similarity. For example, if we take the sentence (or word) X and we calculate its similarity with the word (1) “*happy*”, and then with the word (2) “*miserable*”. If X is more similar to (1) than to (2), that would suggest that X has a higher positive sentiment than negative. To apply the method, we need lists of seed terms with strong semantic orientation (i.e. positive and negative seed lists) but lack sensitivity to context [7] (e.g. good, bad), to use as a reference of sentiment polarity and compare the sentences to them. Pre-existing sentiment lexicons (e.g. VADER, LabMT, etc.) make a great choice to use as seed terms, and can be used whole or partially. For the similarity calculation, Htait et al [4] used a Word2Vec word embedding model. However, Word2Vec lacks the ability to distinguish between the use of a same word in different contexts (e.g. “*bank*” is a riverside or a financial institution) or to understand negations or irony (e.g. “*not happy*”, or “*it was really good, NOT!*”). To solve these problems, we employ neural word embeddings through pre-trained language models (e.g. BERT, etc.) as they can capture the semantics of the entire sentence – which can then be compared to the embeddings of the seed terms. Our SIS method consists of the following components (see Figure 1):

- **Lexicon for the positive and negative seed word lists:** Existing validated lexicons from VADER and LabMT can be used, or a custom specific lexicon can be imported.
- **Neural Transformer for Word Embeddings for the lower dimensional representation of words and sentences:** Using HuggingFace’s Transformer library, different language model can be used directly (e.g. *bert-base-nli-mean-tokens*). Alternatively, such models can be fine tuned specifically to the corpus at hand.
- **Similarity Metric to compute the distance between seed words and target sentences:** So far two similarity metrics have been included: cosine and euclidean distance.
- **Weighting Seed Terms:** To amplify (or not) positive and negative terms, the intensity score of seed terms can be used in conjunction with the similarity score to boost/reduce their impact on the final score.
- **Aggregating function to combine the similarity scores:** To combine the scores of similarity with the terms of the seed lists, an aggregating function is employed: maximum, sum or average. The addition of other similarity metrics or aggregation functions can be easily added in through sub-classes.

Since large seed lists can be computationally expensive, the number of most positive and most negative seed terms from the lexicon can be set. From experiments on three of SemEval’s test collections: SemEval-2016 General English (SE16-GE: 2,999 phrases), SemEval-2016 Mixed Polarity (SE16-MP: 1,269

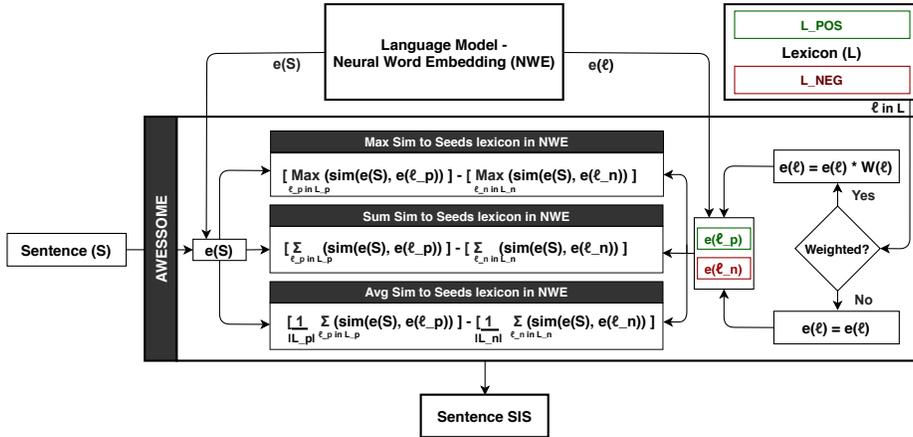


Fig. 1. The AWESSOME framework for SIS consists of three components: a kernel function (e.g. Max, Avg), a neural word embedding and a seed lexicon.

phrases) [5], and SemEval-2018 Task1 (SE18-Vreg: 937 tweets) [8], we found that approximately 500-1000 terms were needed in order to obtain the highest correlations with human annotated intensity scores. While we do not have space to report the details here, we found that the best configurations were by using **600** seed terms from **VADER**'s lexicon, with **BERT** pre-trained language model, while combining the **cosine** similarity scores using the **average** function. Table 1 presents a sample of our experiments results, with the following evaluation measures for correlation coefficient: Kendall's rank (SemEval-2016), and Pearson (SemEval-2018). AWESSOME is open source, freely available from GitHub ¹ and can be directly installed via *pip install awessome*.

Table 1. Sample Results: Correlation between human-annotations and predicted SIS.

Method	SE16-GE	SE16-MP	SE18-Vreg
Supervised -BERT+Linear Classifier	0.502	0.475	0.645
Unsupervised -VADER Module	0.586	0.365	0.517
AWESSOME(AVG,BERT,VADER-600)	0.631	0.563	0.718
AWESSOME(AVG,BERT,LabMT-600)	0.587	0.479	0.642

3 Summary

In this paper, we presented a word embedding-based sentiment scoring framework. The framework combined past works so that pre-existing lexicons and word embeddings can be used to predict sentiment intensity, with no required training, and while providing SIS for words, phrases, and sentences. In another work, we have shown that the performance of the presented SIS methods is comparable or better than existing methods. Our Python 3 toolkit is open source and freely available from GitHub ¹.

Acknowledgement

Cumulative Revelations of Personal Data This project is supported by the UKRI's EPSRC under Grant Numbers: EP/R033854/1.

References

1. Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M.: Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one* **6**(12), e26752 (2011)
2. Gilbert, C., Hutto, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: *ICWSM-14*. vol. 81, p. 82 (2014)
3. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* **1**(12), 2009 (2009)
4. Htaït, A., Fournier, S., Bellot, P., Azzopardi, L., Pasi, G.: Using sentiment analysis for pseudo-relevance feedback in social book search. In: *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. pp. 29–32 (2020)
5. Kiritchenko, S., Mohammad, S., Salameh, M.: Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In: *Proceedings of the 10th international workshop on semantic evaluation (SEMEVAL-2016)*. pp. 42–51 (2016)
6. Liu, B.: Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* **5**(1), 1–167 (2012)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. *ICLR Workshop* (2013)
8. Mohammad, S.M., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: Semeval-2018 Task 1: Affect in tweets. In: *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, LA, USA (2018)
9. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates **71**(2001), 2001 (2001)
10. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. *Journal of the American society for information science and technology* **61**(12), 2544–2558 (2010)
11. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based lstm for aspect-level sentiment classification. In: *Proceedings of the 2016 conference on empirical methods in natural language processing*. pp. 606–615 (2016)