interrupt-driven@SMM4H'24: Relevance-weighted Sentiment Analysis of Reddit Posts

Jessica Elliott and **Roland Elliott**
Independent researchers
send-interrupt@jessicajeanne.co.za
and elliott.roland@gmail.com

Abstract

This paper describes our approach to Task 3 of the Social Media Mining for Health 2024 (SMM4H'24) shared tasks. The objective of the task was to classify the sentiment of social media posts, taken from the social anxiety subreddit, with reference to the outdoors, as positive, negative, neutral, or unrelated. We classified posts using a relevance-weighted sentiment analysis, which scored poorly, at 0.45 accuracy on the test set and 0.396 accuracy on the evaluation set. We consider what factors contributed to these low scores, and what alternatives could yield improvements, namely: improved data cleaning, a sentiment analyzer trained on a more suitable data set, improved sentiment heuristics, and a more involved relevance-weighting.

1 Introduction

Data was taken from the social anxiety subreddit where posts mentioned particular keywords relating to the outdoors. The task was to classify each text as either talking about the outdoors positively with regards to social anxiety, negatively, neutrally, or unrelated to the outdoors or anxiety (Xu et al.). For example this post: "Going for a walk in the rain can be really nice and refreshing so long as I'm wrapped up tight and dry in it." should be classified as positive; "I felt this. One thing I'm always afraid of when going outside is to meet people that I know, especially those of my age." should be classified as negative; "Yeah my eyes are very sensitive when the wind hits or the sun is out." should be classified as neutral; and "..... Run like da wind" would be unrelated.

The approach we took was to classify sentences based on their sentiment (Section 2.2), weight each sentence based on its relevance (Section 2.3), and build a classification for the post based on the average score of each sentence (Section 2.4). We pursued such an approach because, if successful, it has the following advantages: firstly that it is computationally light, requiring minimal resources to be able to run; secondly it is intuitive, making it easy to explain the reasoning for any given classification it makes; and finally it is decomposable, so that each part can be iterated on and improved upon.

2 System description

Here we describe the steps we took to build our relevance-weighted sentiment analysis model.

2.1 Data cleaning and tokenization

Because the data was sourced from social media posts, preprocessing was necessary in order to make the data consistent and well-formed, so as to avoid inadvertent degradation of the sentiment and relevance models.

For example, a common feature in posts was the use of "ill" instead of "I'll", which would artificially lower the sentiment of the sentence. Some posts would add unnecessary whitespace in the middle of words, such as "was n't" rather than "wasn't", which prevented the sentiment analysis from properly identifying negations. And posts used different apostrophe characters, such as "’" and "'", which were not all recognized by the models. We therefore implemented an initial data cleaning preprocessing step to correct these issues. We manually determined the possible cases and then ran detection using regex and string manipulation to transform the data into what we wanted. Some legitimate cases could be transformed erroneously, for example in the case of "ill", but we took this as an acceptable error considering the prevalence of the usage where the poster meant "I'll". Further work could be done on word grammatical analysis to reduce this possibility of error and to test if that would improve the results.

Another common feature of the posts was very long sentences, or sentences not delineated by the

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 98–100

usual punctuation. Since our approach relied upon the tokenization of sentences, we therefore introduced a second preprocessing step, choosing a cut-off length of 40 words and ensuring sentences were no more than that number of words long. After scanning the data we determined that the vast majority of valid sentences were within 40 words or less, which is why we chose this value. When sentences were split in order to achieve this, we included an overlap of 5 words on each side, in order to preserve some of the sentiment-laden context around the 40-word split point. These values seemed to work better than parsing the posts without splitting the sentences, but further work could be done on either using natural language processing to try and determine sentence breaks, or testing different cut-off lengths and overlap windows to see what values achieve the best results.

2.2 Sentiment analysis

The sentiment of each sentence was calculated using the VADER sentiment analyzer (Hutto and Gilbert, 2014), which was pre-trained on the standard VADER lexicon which comes with the NLTK. This analyzer was pre-trained on Twitter (now called X) posts, and returns a sentiment analysis score between -1 (entirely negative) and 1 (entirely positive) for each sentence. The analyzer works by first assigning a default sentiment to each word, which it then modifies based on various heuristics — punctuation, capitalization, degree modification, contrastive conjunctions, and polarity flipping.

We used VADER because its training data is also social media posts, which use informal language and slang terms. However, we identified the following shortcomings when applying it to the current task. First, Twitter is a shorter-form social media platform than Reddit, the analyzer is not therefore designed to deal with the problem of long sentences (discussed in Section 2.1). Second, the analyzer is trained on general sentiment, and would have benefited from a finer-grained contextual training aimed specifically at social anxiety. For example, "staring at me" is considered neutral to the analyzer, but in the context of social anxiety this is something negative. Third, the heuristics introduced to handle negation are not rich enough to capture more nuanced forms of negation found in the longer-form text. For example, offers of advice can use negative words when describing a situation ("If you find walking in public *difficult*, try biking") despite offering a positive sentiment for the advised alternative (biking).

2.3 Relevance analysis

After testing various linear models, we determined that the Passive-Aggressive Regressor (Crammer et al., 2006) performed the best when computing the relevance of sentences. Support for these regressors is included in the SciKit-Learn Python library. We trained our regressor as follows. First, we applied a term frequency-inverse document frequency transformation on trigrams of the data (ignoring common English stop words). Second, we provided as labeled data the sentences of the training posts which contained one or more keywords, labeled as relevant (1) or irrelevant (0) based on whether the post as a whole was related (positive, neutral, or negative) or not (unrelated) as given by the task data. The resulting regressor returns a relevance score between 0 and 1 for any given text.

When classifying test data, the regressor is applied to the post as a whole to determine its relevance. If its relevance score is less than 0.25, then the post is classified as unrelated. Otherwise, it passes to the next stage, to which we now turn.

2.4 Relevance-weighted sentiment scoring

In order to calculate the overall sentiment of a post, we combined the sentiment analyzer s and the relevance regressor r described in the previous two sections. For each sentence t_i, we define the adjusted relevance R with the recurrence relation:

$$R(t_i) = \max\{r(t_i), 0.9 \cdot R(t_{i-1})\},$$

with $R(t_0) = rel(t_0)$ as the base case. This adjusted relevance allows us to model the fact that relevance can be inherited from previous sentences, with a topic being broached in one sentence and continued in later sentences without the same words necessarily appearing in them. The decay factor of 0.9 captures the intuition that the further from the original sentence we get, the less likely it is still what is being spoken about.

We use R to weight the sentiment of each sentence in the post, and then take the average of these weighted scores. As with r, R returns a relevance score between 0 and 1, so that this relevance-weighted sentiment score is calculated as:

$$\frac{\sum_{i=0}^{n} R(t_i) \cdot s(t_i)}{n}.$$

3 Evaluation

The resulting scores from our model were low. On the validation data, this approach scored as follows:

Label	Accuracy
Unrelated (0)	0.55
Positive (1):	0.33
Neutral (2)	0.18
Negative (3)	0.55
Total	0.45

And on the evaluation data, it scored as follows:

Score	Value
F1 score	0.358
Precision	0.365
Recall	0.411
Accuracy	0.396

As discussed in each section, the system and tools chosen did not fit well with the type of data provided. We consider the approach to be one that could work if the suggested improvements are made, such as improved data cleaning, a sentiment analyzer trained on a more suitable data set, improved sentiment heuristics, and a more involved relevance-weighting. As a comparison with other methods, here are the mean and median results of all the teams:

	Mean	Median
F1 score	0.5186	0.5795
Precision	0.5649	0.63
Recall	0.5379	0.5885
Accuracy	0.5746	0.627

4 Conclusion

We use a relevance-weight sentiment analysis approach for classifying the sentiment of Reddit posts with respect to social anxiety and the outdoors. Even though our particular implementation produced poor results, there are opportunities for improvement in almost every stage of the approach: improved data cleaning, a sentiment analyzer trained on a more suitable data set, improved sentiment heuristics, and a more involved relevance-weighting each improve the overall performance of such an approach.

References

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer, and Manfred K Warmuth. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(3).

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

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. Overview of the 9th social media mining for health applications (#SMM4H) shared tasks at ACL 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks", month = Aug, year =*.