Strong Winds and Active Volcanoes: Unsupervised Methods for Linguistic Analysis of the Weather

Keywords: Sentiment Analysis, NLP, Extreme Weather, Social Media, Social Sensing

Extended Abstract

Social sensing is the systematic collection and investigation of social media data to improve our understanding of real-world events. Despite its widespread use in detecting and locating extreme weather events for over 10 years, social sensing research often stops at detection, while quantitatively measuring public mood and the real-world impacts of extreme weather using online media data remains challenging. Accurate identification of emotive posts on social media would enable better assessment of major weather events, but current sentiment analysis techniques are either limited in accuracy or require significant time and resources to train. We demonstrate an unsupervised approach for identifying the significant linguistic variations between weather events, before utilising these differences to improve current sentiment analysis techniques. Additionally, we show that language alone can indicate both wind speeds and ambient temperatures, with particular words correlated with different speeds and temperatures.

Whilst general-purpose sentiment analysis approaches tend to perform well on average, they often struggle when applied to specific domains. For example, VaderSentiment (Vader) (Hutto and Gilbert, 2014), a popular social media optimised rule-based dictionary approach, classifies, “Help me, there is a very strong storm #alert!!” as very positive, due to Vader assigning positive scores to “help”, “strong”, and “alert”. Manually updating the sentiment lexicon for each new domain is impractical, as it requires significant human effort to understand the contextual variations of each corpus. Another approach is the supervised training of a novel classifier for a given domain. However, these methods are costly, requiring thousands of sample messages to be manually labelled.

To address these challenges, we use an unsupervised approach based on SocialSent, a technique developed by Hamilton et al. (2016), which can create a domain-specific sentiment lexicon from a small set of positive and negative seed words as input. SocialSent involves forming a lexical graph based on semantic relationships between vectorised words in a dataset. Label propagation from the location of a small number of known positive and negative seed words within the graph is carried out, measuring each word’s proximity to the positive and negative seeds. The sentiment polarity for every word in the graph is then calculated using each word’s respective positive and negative proximity. Our approach is to construct a domain-specific dictionary using SocialSent, and then substitute the values into Vader. This is to make use of Vader’s built-in mutation rules (negation, boosting, etc.), for instance, Vader assigns “that is very BAD!” as more negative than “that is bad”. To validate how well the SocialSent derived polarities agree with the manually assigned Vader polarities, we applied it to 14 million random tweets from Great Britain in 2020. The resulting polarities showed a strong correlation ($r = 0.86$, $P < 0.001$) with the pre-existing sentiment lexicon used by Vader.

To demonstrate the importance of domain-level sentiment lexicons, we compared the general-corpus polarity scores from Vader to specific-corpus polarity scores assigned by SocialSent. First, we obtained manually filtered weather tweets collected by Asiaee T. et al. (2012) as part of the “Dialogue Earth” project. These were separated and filtered into three weather conditions: ‘heat’, ‘wind/storm’, and ‘cold’. The union of these three subsets was included as a fourth dataset. Figure 1 shows the percentile rank position of each distinct word for its polarity score from both Vader and SocialSent-derived polarity. The highlighted words are the 8 words with the greatest rank difference. The results are largely intuitive, for instance, Vader assigns warmth as a positive word, whereas it is negative in the heat dataset. The same can be seen for “alert”, “help”, “strong”, “care”, and “special” in the other plots. Two interesting examples are “NH”, and “TX”, both of which are positive in the Vader dictionary, however, in the dataset they are used as abbreviations when discussing extreme weather in “New Hampshire” and “Texas”. Finally, many of the other positive Vader words which are returned as negative for SocialSent (“lmao”, “lol”, “haha”, “funny”), when investigated, showed up in sarcastic tweets, which Vader would incorrectly classify as strongly positive. Upon substituting the filtered polarities into Vader, the tweets were reclassified, with 14% of the dataset switching polarity entirely.

The SocialSent method can also be used to map words to scales other than sentiment, by initiating the algorithm with alternate sets of seed words. Here we use this property to derive linguistic scales for wind speeds and temperatures, finding a clear relationship between the derived polarities (assigned from these otherwise-arbitrary linguistic scales) and the true wind and temperature values associated with each tweet. The dataset investigated for this was a 2020 geolocated set of 14 million tweets from Great Britain. To determine the true temperature’s position on the SocialSent-derived scale, we encoded and appended the rounded temperature of each tweet to the tweet text, using the time and location of each tweet in the dataset, alongside gridded temporal weather data. SocialSent was then applied using a selection of hot and cold words and emojis as the seed set, creating a hot-cold linguistic spectrum. This was then repeated for wind speed using calm and stormy seed words to create a calm-stormy spectrum.

Figure 2 shows the scores of the encoded temperatures and wind speeds within the derived lexicon, compared to their decoded temperature and wind speed. The strong relationship between the SocialSent temperature/wind polarities and the observed temperature/wind speed suggests that linguistic variation in non-weather-related data is affected by weather conditions. Additionally, there is a social saturation effect, where after a certain temperature (∼20°C) and wind speed (∼8.5 m/s), the language used alongside the more extreme conditions plateaus. This highlights the importance of using more than just weather observations when trying to understand how people are impacted by extreme weather.

Whilst this methodology is not limited to weather events, this study demonstrates the importance of domain-specific sentiment analysis and its potential for advancing social sensing. The research emphasises the versatility and efficiency of SocialSent in overcoming limitations of current supervised approaches. Not only can it be used for improving understanding of sentiment, but also for physical scales that are difficult to manually rank/categorise, such as the ‘windiness’ or ‘temperature’ of tweets. This approach presents an opportunity for further development in our understanding of language and its relationship to extreme weather, as well as the potential for more accurate and efficient social sensing in a variety of domains.

References

Asiaee T., A., Tepper, M., Banerjee, A., and Sapiro, G. (2012). If you are happy and you know it... tweet. Association for Computing Machinery.

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. *arXiv:1606.02820 [cs]*.

Hutto, C. and Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *ICWSM*.

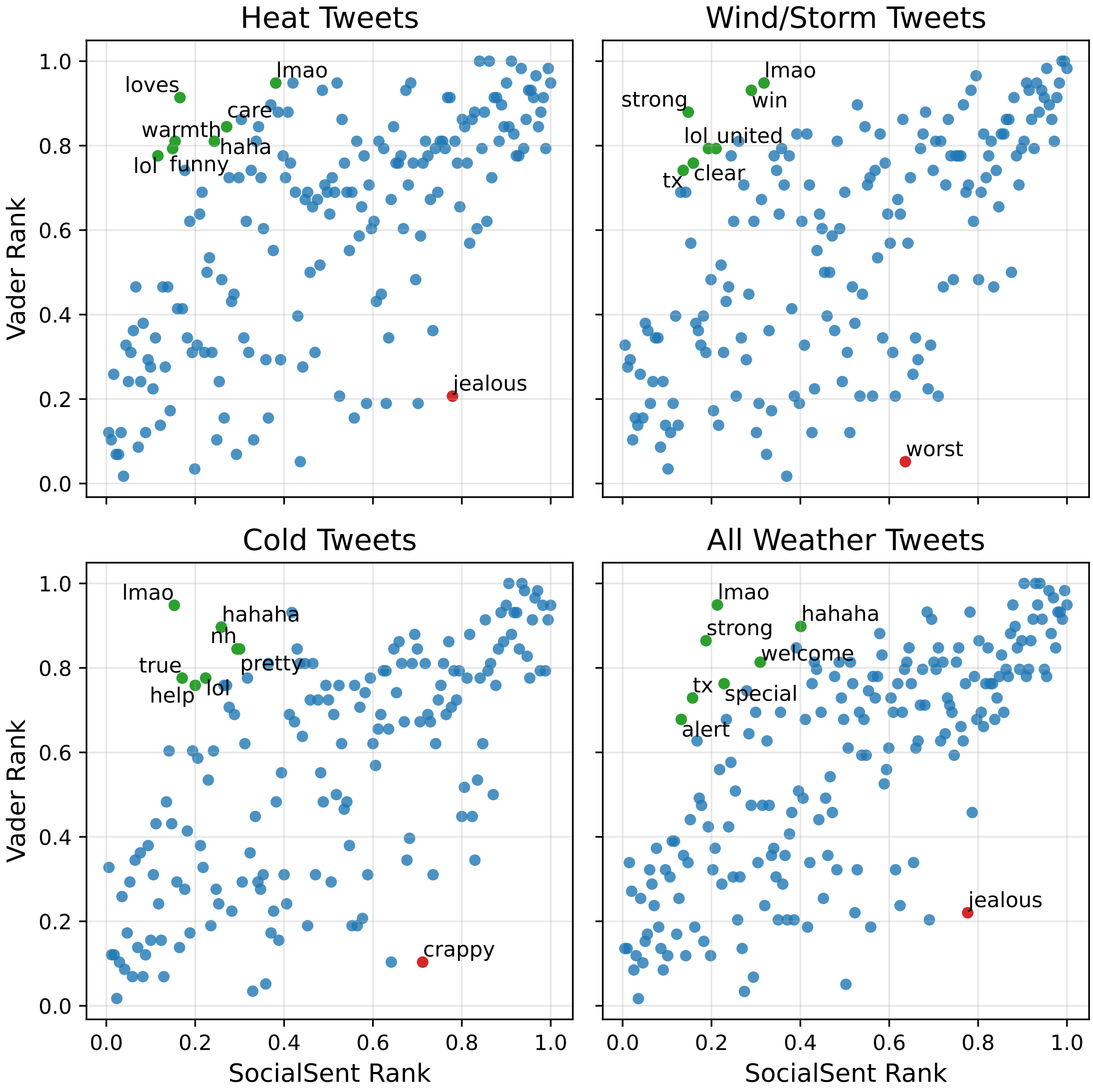

Figure 1: Percentile rank comparison of VaderSentiment lexicon and SocialSent-derived polarities for the different weather event datasets. The 8 words with the greatest rank difference have been highlighted.

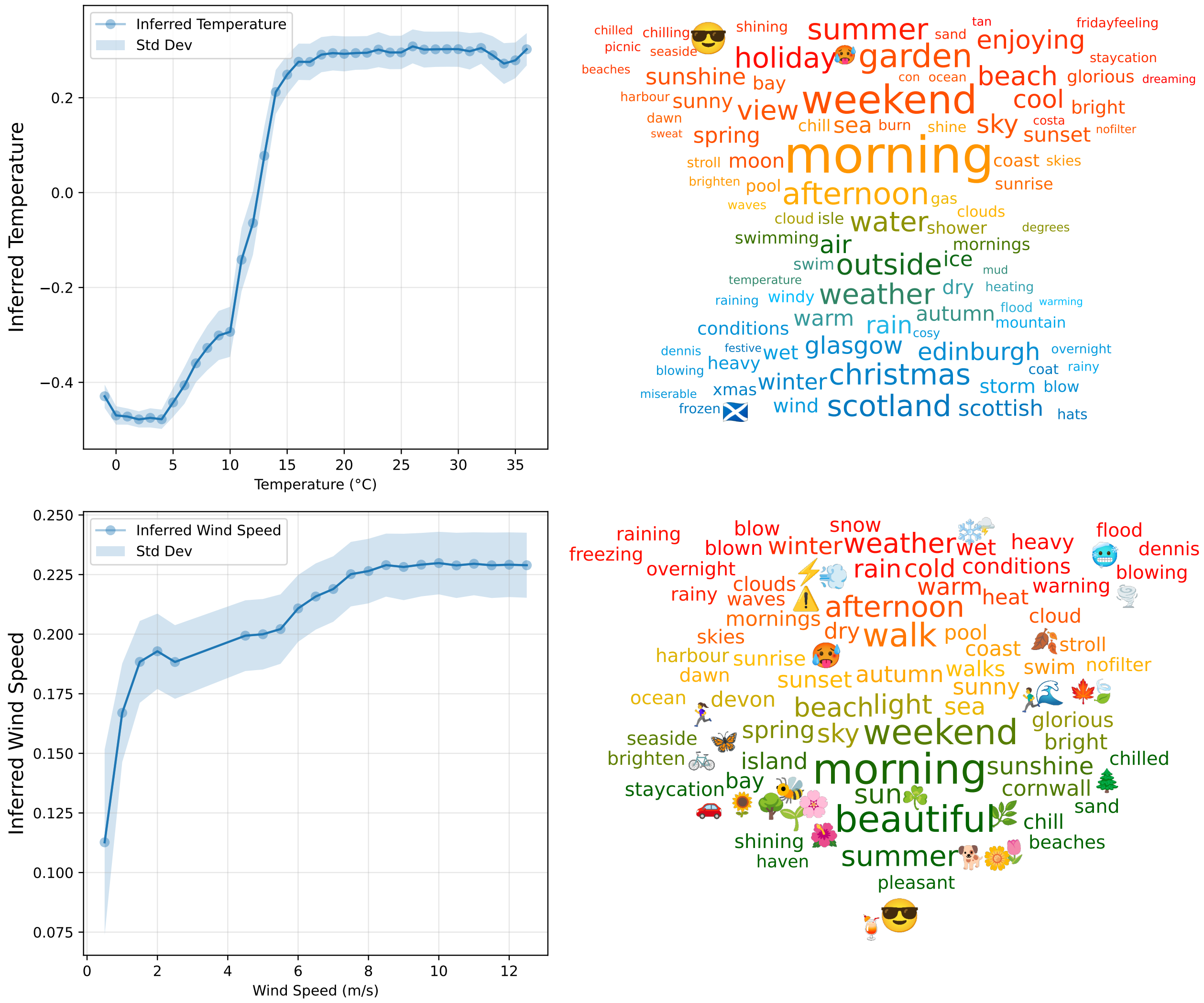

Figure 2: Scores of the encoded temperatures and wind speeds within the derived lexicon, compared to their decoded temperature and wind speed. The wordclouds show how the language varies along the derived scales.