Sentiment Analysis of the ECO Hotel

Christian R Chasteau

Abstract

The dataset used is based upon comments from guest about their experience when staying at ECO Hotel [1].

Can we select a sentiment analysis technique which will lend itself to being incorporated into a solution for a business user.

Vader [2], was chosen as a Lexicon rule based method specifically built for microblog social media feeds. This produced the expected polarity for a single sentence, paragraphs required an algorithm to extract the lowest negative value. The range used for selecting the negative sentiment was altered from the default to taking into account of results found during the analysis.

Motivation

It only takes one wrong decision and a business can find themselves with their reputation under threat; which even the best PR will have difficulty resolving, having a negative impact on their business [6]. This may be the worst case scenario, but reputational risk costs [7], and can cost dearly. With the advent of social media, a simple negative comment can spread with limited recourse.

Just as firewalls are a necessary asset in protecting a company's network, so we can say is understanding and managing sentiment. Dealing with primarily negative sentiment quickly, protects the company, lessening the impact to their reputation.

Any method of classifying sentiment from social media, would help to mitigate such a PR calamity. From the multiple methods of sentiment analysis[3] we just need one. Business could probably ask a data scientist to pick one method of sentiment analysis and present a baseline solution that could be productionised. It should be easy to implement. Requires minimal maintenance when in production and serve a early warning to any negative sentiment.

Dataset

This dataset includes Online Textual Reviews from both online (e.g., TripAdvisor) and offline (e.g., Guests' book) sources from the Areias do Seixo Eco-Resort.

The CSV holds one cell per review, which could be made up of a single or multiple sentences, except the first row, which is the header, 401 attributes. All the reviews were collected between January and August of 2015.

The dataset can be found here: https://archive.ics.uci.edu/ml/datasets/Eco-hotel

Acknowledgement for use of this data will be found in the Acknowledgement section

Data Preparation and Cleaning

At a high-level, what did you need to do to prepare the data for analysis? Describe what problems, if any, did you encounter with the dataset?

The data did not contain any hyperlinks or any emoticons. I still had to remove records which contained only spaces, trimming leading and trailing spaces as well. This dataset includes utf8 characters and therefore care must be taken when converting the data during any cleansing.

Research Question(s)

From the motivation section, with modifications specific for this dataset:

“A Hotelier, in this case could probably ask a data scientist to pick one method of sentiment analysis and present a potential solution that could be productionised. It should be **easy to implement**. Requires **minimal maintenance** when in production and serve a early warning to any negative sentiment.”

Following on from this, our simplified research question will be.

Can you choose method of sentiment analysis applying to the hotel reviews dataset, determining if it will meet the given business requirement from above. That is, the ability to show any negative sentiment.

Methods

You'll need to back your claim with hard evidence

When researching methods for sentiment analysis one approach was mentioned a number of times: Vader [2]. I eventually found a paper[3] were comparisons for sentiment analysis had been carried out and these theses[4][5] were Vader was used.

Vader is appropriate as it combines a dictionary, which maps lexical features to emotion intensity, and five simple heuristics, which encode how contextual elements increment, decrement, or negate the sentiment of text.
Sentiment can be categorical – such as {negative, neutral, positive} – or it can be numerical – like a range of intensities or scores. Lexical approaches look at the sentiment category or score of each word in the sentence and decide what the sentiment category or score of the whole sentence is. The power of lexical approaches lies in the fact that we do not need to train a model using labeled data, since we have everything we need to assess the sentiment of sentences in the dictionary of emotions.

Methods contd.

The returned result of any analysis the polarity score is as shown below

```
{'compound': -0.3252, 'neg': 0.437, 'neu': 0.563, 'pos': 0.0}
```

Compound is valence summation of each word in the lexicon and adjusted. Values range from -1 most negative to 1, most positive. It will be the compound[8] value for a sentence that is used.

All reviews were subjected to tokenisation of a sentence prior to analysis. Although the reviews will have multiple sentences, sentence tokenization will produce a list of tokenized sentences.

The defaults for classification are

- Compound score >= 0.5 being positive sentiment
- Compound score <= -0.5 being negative sentiment
- >=0.5 to <= -0.5 as neutral sentiment

Vader is availabe as part of the Python library https://pypi.python.org/pypi/vaderSentiment or the NLTK library http://www.nltk.org/_modules/nltk/sentiment/vader.html

Findings

The focus of the findings relate to the value of the negative sentiment, whether there is a single sentence or paragraph. Finding how to get the correct value is important as we cannot answer the research question.

First we'll find out what exactly Vader makes of the following sentences

```
In [553]: compoundSentiment("I do not think such situations are acceptable.")

Out[553]: (0.3182,
           ['I do not think such situations are acceptable.'],
           {'compound': 0.3182, 'neg': 0.0, 'neu': 0.723, 'pos': 0.277},
           array([ 0.3182]))

In [554]: compoundSentiment("It wasn't a nice meal")

Out[554]: (-0.3252,
           ["It wasn't a nice meal"],
           {'compound': -0.3252, 'neg': 0.437, 'neu': 0.563, 'pos': 0.0},
           array([-0.3252]))
```

Findings contd.

For paragraphs, the best approach would be to collate all compounds and find the negative value, if there is one

```
In [556]: compoundSentiment("That doesn't mean their are better ones out there")

Out[556]: (0.4404,
           ["That doesn't mean their are better ones out there"],
           {'compound': 0.4404, 'neg': 0.0, 'neu': 0.734, 'pos': 0.266},
           array([ 0.4404]))

In [557]: compoundSentiment("I think that Vader is the worst analysis tool ever. That doesn't mean their are better ones out ther

Out[557]: (-0.6249,
           ['I think that Vader is the worst analysis tool ever.',
            "That doesn't mean their are better ones out there"],
           {'compound': 0.4404, 'neg': 0.0, 'neu': 0.734, 'pos': 0.266},
           array([-0.6249,  0.4404]))
```

From above we see that the array contains both sentences compound values and the function returns the correct negative value of -0.6249. This is correct and is exactly as we need from our research question.

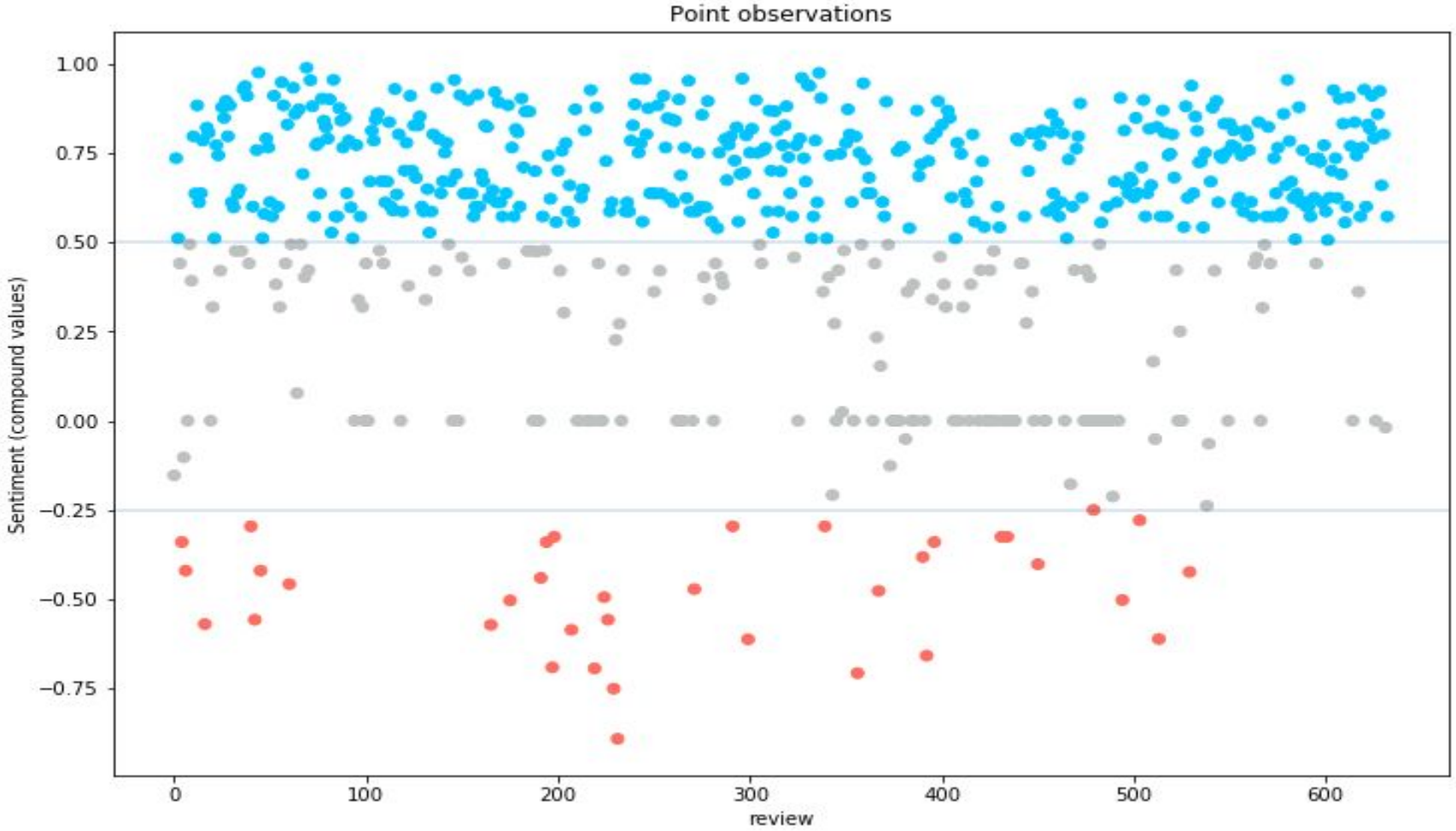
Findings contd.

In the final slide we change the starting point at which negative sentiment is classified. Originally it was per the default, which was -0.5

This was changed because from the first slide, the result showed a negative value of -0.3252. So now we classify a negative sentiment starting from -0.25

Also I’ve incorporated the change as shown in the previous slide to improve classification of paragraphs in a review.

The meets the research criteria, ensuring we don’t miss any negative sentiment.


Point observations
Sentiment (compound values)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0
100
200
300
400
500
600
review

Limitations

The ECO Hotel is located in Portugal. The dataset was limited to the English language and from predominantly a single source, Trip Advisor. Other sources, excluding those from the hotel, could have been incorporated. This would mean adding a second language to a single dataset or a separate Portuguese dataset.

Conclusions

Report your overall conclusions, preferably a conclusion per research question

Can you choose method of sentiment analysis applying to the hotel reviews dataset, determining if it will meet given business requirement from above.

1. When paragraphs were analysed, the original approach was to average out the compound score which lead to inaccurate sentiment. By collating them together, regardless as to whether it is for a single sentence or paragraph, returning the compound score with the most negative value gives a more accurate sentiment for that review.
2. The Vader documentation points to using -0.5 as the starting point at which negative sentiment should be classified. I found that during the analysis, a single negative sentence from the review could produce a negative value closer to -0.3. So I adjusted the value at which the negative sentiment would apply to -0.25 instead.
3. Modifications as shown from items 1 and 2 you could implement a warning system on negative sentiment using Vader.

Acknowledgements

No feedback was asked for or given for this project.

The data set “Eco-hotel Data Set” was retrieved from https://archive.ics.uci.edu/ml/datasets/Eco-hotel

Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment Classification of Consumer Generated Online Reviews Using Topic Modeling. Journal of Hospitality Marketing & Management, DOI: 10.1080/19368623.2017.1310075.

I would also like to acknowledge these two theses which I found to be of great help, their knowledge and insight help me immensely in this project.

- Cambero, Angel, "A Comparative Study of Twitter Sentiment Analysis Methods for Live Applications" (2016). Thesis. Rochester Institute of Technology http://scholarworks.rit.edu/cgi/viewcontent.cgi?article=10367&context=theses
- Sentiment Classification in Social Media https://www.diva-portal.org/smash/get/diva2:930520/FULLTEXT01.pdf

References

The work was carried out with the help of the references below.

[1] Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment Classification of Consumer Generated Online Reviews Using Topic Modeling. Journal of Hospitality Marketing & Management, DOI: 10.1080/19368623.2017.1310075. https://archive.ics.uci.edu/ml/datasets/Eco-hotel

[2] VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text (by C.J. Hutto and Eric Gilbert). Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014. http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

References contd

[3] SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods https://arxiv.org/pdf/1512.01818.pdf , https://link.springer.com/article/10.1140/epjds/s13688-016-0085-1/fulltext.html

[4] Cambero, Angel, "A Comparative Study of Twitter Sentiment Analysis Methods for Live Applications" (2016). Thesis. Rochester Institute of Technology http://scholarworks.rit.edu/cgi/viewcontent.cgi?article=10367&context=theses

[5] Sentiment Classification in Social Media

https://www.diva-portal.org/smash/get/diva2:930520/FULLTEXT01.pdf

References contd.

[6] Harvard Business Review (Vanitha Swaminathan, Suyun Mah; 2016-09-02) https://hbr.org/2016/09/what-100000-tweets-about-the-volkswagen-scandal-tell-us-about-angry-customers

[7] https://en.wikipedia.org/wiki/Reputational_risk

[8] Explanation of compound polarity score https://github.com/cjhutto/vaderSentiment#about-the-scoring