Developing an improved framework for correctly identifying and understanding the effects of sarcasm on polarity scores

DSM060 – Coursework 2

Word count: 2919

1. **Introduction and motivation**

Polarity scores are a well-established metric in the world of natural language processing for understanding the sentiment of a given word, sentence, or body of text (C. Bhadane, et al, 2015). These scores can be used to better help a restaurant owner capture the general sentiment customers have towards their business through reviews, or to help create a targeted marketing campaign strategy for a given product, or even to understand the general opinion of the masses on topics such as politics, or celebrity gossip. All the before mentioned applications for polarity scores are cursed with a major problem; it is exceptionally hard to capture the sentiment, and therefore accurate polarity scores for informal user-written text, which often includes implicit language, such as sarcasm (M. Tubishat, et al, 2018). The importance of understanding implicit language for polarity score calculations can be seen in Maynard’s 2014 study where a simple model for identifying sarcasm was created. Using this model they identified a 50% increase in polarity accuracy when applied to a set of tweets that had been identified as containing sarcasm, and when applied to a completely unseen set of tweets polarity accuracies were increased by 11% as compared to a traditional polarity model that did not consider sarcasm. Clearly, to truly understand the benefits of polarity scores, we must first ensure they are calculated correctly, to do this sarcasm must be correctly accounted for.

**2.0 Objectives**

As previously mentioned, there have been many significant advancements in the field of sarcasm detection (Onan & Toçoğlu, 2021; Carvalho & Guedes, 2020; Martino et al, 2020; Sundararajan & Palanisamy, 2020), but very few advancements in the use of detected sarcasm to improve the accuracy of polarity scores (CT, 2022). In this project we aim to change this by developing a machine learning framework that can accurately account for sarcasm in sentiment analysis, resulting in a fairer and more accurate representation of the sentiment being displayed in a body of text. This will lead to more significant insights being drawn from text-based sentiment analysis. The framework we introduce also has generalizability in mind, meaning it should be useful for all kinds of text formats, something that is not seen in much of the previous works in the field (Onan & Toçoğlu, 2021; CT, 2022).

To achieve our overall goal of developing a framework for the accurate understanding of sarcasm sentiment, we will need to pass four goals, which correspond to the below 4 layers in fig. 1.

Fig. 1 An overview of the 4 layers of our polarity model

**Goal 1**: **Create a baseline sentiment analysis model.**

The easiest, but potentially most important step in this project is producing a good-baseline sentiment analysis model. Creating a generic sentiment analysis model is outside the objectives of this project and so we will use a pre-built, open-source model found in the literature. L. Mathew & V. R. Bindu (2020) give an in-depth overview of several cutting-edge, pre-built models for sentiment analysis and will be the primary resource used for implementing such a model.

**Goal 2**: **Create an extra layer to the sentiment analysis model created in Goal 1**

After we have our baseline model, we will aim to add an additional machine learning layer to the framework that will form the core of this project. The purpose of this layer, which we will refer to as the *hierarchical anomaly detection layer* (HADL), is to highlight specific text-structures within the hierarchy of a piece of text, such as words, sentences, or paragraphs text that have *anomalous* polarity scores. Anomalous polarity scores are defined as being those which are significantly different relative to the text-structures surrounding it. The aim of this layer is to detect the specific areas of text where sarcasm takes place, and from there we can adjust polarities according to the type of text that is highlighted. Fig. 2 shows an example hierarchical text-structure for an article. Two anomalous polarity scores can be seen, where one corresponds to a particular word, and the other to a particular sentence. They can be said to be anomalous since their polarity scores are significantly different to both their sister nodes, and parent node.

Waterfall chart

Description automatically generated with medium confidence

Fig 2. An example hierarchical text-structure with 4 layers, showing example polarity scores at each stage. The two structures highlighted in red are example nodes that would be highlighted in layer-2 of our framework. The reason being, their polarities are significantly different to their parent node, and sister nodes.

Using a simple example, we can further explain the above figure, and show the workings of this layer in the framework. Take the sentence:

“My neighbour was shouting all night and I slept horribly, great start.”

In this example, the sentence’s polarity score would be slightly negative, there are a few instances of very negative words being used, and a single instance of a very positive bigram phrase, “great start”. Our model would identify that “great start” is an anomaly when comparing its polarity with that of the other phrases in the sentence, and the sentence itself, thus our model would highlight this node as shown above. It is important to note that, despite looking like two words, great start would be counted as a single phrase in our model since it is a common bigram. This same logic would apply at any level in the hierarchy.

**Goal 3: Identify the type of anomaly that was highlighted by HADL**

The next major step in the project is to identify and classify the type of anomaly highlighted in the second layer of the framework, which will give us an idea of how we should transform its polarity. To do this, we will take the text-structure highlighted as an anomaly and search through it for different anomaly types. The main anomaly types we will be dealing with are:

* **Negative sarcasm.** Negative sarcasm is simply when a word or phrase is used to mean the polar opposite of its actual definition, this kind of sarcasm can be identified very easily by looking for negation words (e.g., not, no-one, nobody).
* **Idioms.** Another simple anomaly. For the English language there are large repositories of idioms that store their intended meaning. We can simply do a search through the relevant text-structure to identify if any idioms are present.
* **Irony**. This will typically be identified in the more granular levels of the hierarchy, if an entire sentence or group of sentences are seen to be anomalous, irony can be assumed.
* **Aggressive Sarcasm**.This will be identified as anomalies that have a target i.e., a person’s name. This will be verified by using Part Of Speech (POS) tagging.

There are several other types of sarcasm that could be added to the above list, but due to the scope and length of the project we will only consider the above.

**Goal 4: Polarity transformation**

To finish the model, we will need to transform the polarities of the text that have been identified and classified as a particular type of sarcasm. As this is an un-supervised task, validating the transformation we make will have to be an iterative process. If available, it would be useful to collaborate with a linguistics expert for their insight into how the different kinds of sarcasm would affect polarity, or more broadly, sentiment. For this project, we will initially make some broad assumptions on how polarity scores should be transformed for each type of sarcasm. Polarity scores will then be compared with those calculated for the re-phrased data for validation. A visual overview of Goals 3 and 4 can be seen in fig 3.

Fig. 3 The pipeline to be implemented in goal 3 and 4, where an anomaly is taken, classified, and transformed.

**5.0 Benefiting stakeholders**

As mentioned previously sentiment analysis has a large impact on a variety of industries, many of which rely on informal, user-written text where implicit language such as sarcasm is very common. This project will greatly improve the accuracy of sentiment analysis, directly affecting and improving the understanding that can be gained from sentiment, and from that, the potential for significant monetary gain. For example, a better understanding of your target audience’s needs through accurate, and representative sentiment analysis can ensure a company correctly identifies the needs of its customers. This can lead to better product design or a more realistic view of your restaurant’s reviews, leading to meaningful and impactful changes towards business improvement. Alternatively, understanding the honest opinion of the masses on a particular political event can help a campaign organiser better understand the good, and bad parts of their campaign strategy. In general, we believe that using sentiment analysis for a more representative analysis of a population can directly improve the performance of a great variety of businesses.

**6.0 Related work**

Despite the importance of sarcasm on sentiment analysis, which has previously been empirically shown by both Maynard (2014) and Karoui et al (2015), there has been limited work in developing models that can make polarity transformations to represent sentiment more accurately. Of the few studies that have been done there are none that are comprehensive or make the attempt to understand the nuances of sarcasm, such as the different kinds described above. As an example, both the Maynard and Karoui studies previously mentioned, use a rather simplistic method for polarity transformation. They either look for negation words, and then flip the polarity for that sentence or phrase, or they look for single indicator words that explicitly denote sarcasm, in which case polarity is again simply flipped. Maynard (2014) also made use of hashtags in tweet data to see when words such as sarcasm, or irony were used. Of course, this is an oversimplification of the problem, and doesn’t not generalise at all to non-tweet data. Despite these huge limitations, a significant positive effect was seen on the polarity scores calculated.

With that in mind, there are many models that have been developed that are exceptional at detecting the presence of sarcasm but have no framework in place to understand its effects on polarity. Sundararajan and Palanisamy (2020) introduce a two-part approach that made use of an ensemble-based feature selection model and a multi-rule-based algorithm for sarcasm detection and classification. Their model saw f1-scores of over 92% on un-seen data. Wen, et al (2022) introduced a novel neural network model that used word-embedding and auxiliary information as enhancements to improve previous sarcasm detection models. They achieved f1-scores of 73%, which is particularly impressive since they used longer-format text data, such as news articles, for training, where all other studies mentioned here used short-text formats from social media. The year prior, Onan and Toçoğlu (2021) also created a novel neural network model that proved to be cutting-edge in the sarcasm detection of twitter text data, with their model consistently seeing f1-scores greater than 83% on a variety of different datasets, although all datasets used, contained exclusively twitter data. Despite the success of all of these studies, almost all of them proved to have very specific use-cases, and in general generalised poorly. More specific to this project, they also ignored the fact that; in general, detecting sarcasm isn’t that useful. Despite sarcasm detection having limited uses, understanding sarcasm and the effects it has on sentiment is highly useful and the focus of this project. For a more comprehensive overview of related work see our previous work (C. Thickett, 2022).

**7.0 Data**

The SemEval Task11 dataset (I. A. Farhu, et al, 2021) contains over 16,000 short-form texts like those that could be seen on a social media platform. All the data has been labelled as either sarcastic, or not sarcastic. These labels were provided by the author of the text itself, since the initial classification of sarcasm was done manually by the author, the validity and accuracy of the labels should be very high. The authors also provide a rephrased body of text that depicts the literal meaning of the sarcastic word, phrase, or sentence. This allows us to compare our transformed polarities calculated from the text containing sarcasm, with the more literal, rephrased text that conveys the same message in a non-sarcastic manner. The dataset can be downloaded for free from their website (I. A. Farhu, et al, 2021) and as such there are no ethical or moral issues surrounding the data. If for any reason the SemEval dataset becomes unavailable or is not fit-for-purpose, twitter has an excellent API that we can utilise to scrape our own sarcastically labelled data. Of course, this would not include any re-phrased data and so will only be used as a last resort.

**8.0 Methods**

**9.0 Work plan**

Chart, waterfall chart

Description automatically generated

Fig. 4 A gnatt chart for the break down of each project goal, and the time they are expected to take.

Fig. 4 displays the full work plan for the project, the key points to note are that at the end of each section, or goal, time is taken to validate the model. This ensures that at each goal end we can be sure that our work thus far is meaningful and heading in the right direction. The first two steps are predicted to take the longest and involve a significant amount of work, for this reason the time allocated to validation for goal 1 and 2 is higher than for goal 3 and 4. This gives us extra time to make iterative changes to the model depending on how it is performing. Where possible, validation and checks of the model will happen alongside development of the model.

**10.0 Risk assessment**

Several potential setbacks are possible within this project, the major risks we foresee happening are listed below:

* **we are unable to re-create a cutting-edge sentiment analysis model for the basis of our model.** 
  + If this is the case, we will simply opt to make a simple Naïve-Bayes model that makes use of word-embedding and tf-idf features. The primary goal of the project is not to create a base sentiment analysis model, so this is a manageable issue that shouldn’t impact the results of the project, although it does mean we will not be able to directly compare our models to the current ones found in the literature.
* **Data imbalances** 
  + We aim to deal with sarcasm differently depending on the type of text-anomaly it belongs to. It is possible that some anomalies will be significantly rarer than others. To solve this, we will under-sample the majority groups to create more balanced classes. If this does not solve the issue, we can look to extract additional data from earlier versions of the SemEval dataset, although they do not contain re-phrased text so its usage will be limited.
* **Issues creating logic for anomaly classifications** 
  + Although the initial ideas are present for how we will classify different anomalies, it is possible that due to limited time some anomalies will prove to be significantly harder to classify than others, if this is the case, we will focus our efforts on a sub-set of the anomalies.

**11.0 Expected Results**

The successful completion of this project will yield a generalizable model that is capable of understanding, and accounting for the effects sarcasm has on sentiment. It will be built on top of an already successful sentiment analysis model, although adding our framework to any model that calculates polarity scores should be quick and easy. We also introduce a novel way of looking at text for the purpose of polarity calculation, namely the HADL. By correctly classifying, and transforming polarities of sarcastic text, we should be able to conclude that correctly accounting for the effects of sarcasm can lead to significantly enhanced sentiment analysis models that will have more use-cases in industry. Due to the presence of our HADL, our results should be well generalised, in fact we would expect to achieve even better results when our model is applied to larger text bases that have more hierarchical layers.

**12.0 Evaluation**

Fig. 5 The evaluation process for our final model where we use the re-phrased SemEval data and our base-sentiment model to calculate target polarities. We then compare the polarity scores calculated using both the base model, and the HADL on the sarcastic text to calculate RMSE and compare the performances of the two models.

As mentioned in the workplan, validation will take place at the end of each goal. In the first step we aim to re-create a baseline sentiment analysis model from the literature. To evaluate the success of this step we will compare its outputs to those seen in the relevant paper. Next, we will look to ensure the anomaly classifications from the addition of HADL makes sense. This step is un-supervised and as such validation will have to be done manually, i.e., by looking at the classifications made and seeing if they make intuitive sense. Although, if this step performed poorly, it will also be seen in the evaluation of the final model. The final, and most important evaluation will take place at the end, where we will use the polarities calculated from the re-phrased text data as our dependant variable. Since we are directly comparing scores, we can use metrics such as RMSE to see how far away we are from a “true” polarity assignment by incorporating sarcasm. This can then be compared to how our base model performs on the sarcastic text data to give an imperial view of the improvements made by incorporating our novel sarcasm model, see fig. 5 for a visual representation of the evaluation process for the final model.

**References**

1. D. Maynard, 2014, ‘Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis’, *European Language Resources Association (ELRA)*, pp. 4238-4243.

2. A. Onan and M. A. Toçoğlu, 2021, ‘A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification’, *IEEE*, vol. 9, pp. 7701-7722

3. C. Thickett, 2022, ‘Sentiment analysis: A look at the effects of sarcasm’, *University of London*, DSM060.

4. F. Carvalho, G. P. Guedes, 2020, ‘TF-IDFC-RF: a novel supervised term weighting scheme’, arXiv preprint.

5. G. D. S. Martino, A. B. Cedeno, H. Wachsmuth, et al, 2020, ‘SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles’, International Committee for Computational Linguistics, pp. 1377-1414.

6. K. Sundararajan, A. Palanisamy, 2020, ‘Multi-Rule Based Ensemble Feature Selection Model for Sarcasm Type Detection in Twitter’, *Computational Intelligence and Neuroscience*, vol. 2020, pp 17.

7. L. Mathew, V. R. Bindu, 2020, ‘A Review of Natural Language Processing Techniques for Sentiment Analysis using Pre-trained Models’, *Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 340-345.

8. I. A. Farhu, S. V. Oprea, S. Wilson, W. Magdy, 2021, SemEval2022, [https://sites.google.com/view/semeval2022-isarcasmeval/home#h.t53li2ejhrh8](about:blank), last accessed: 28/08/2022

9. C. Bhadane, H. Dalal, H. Doshi, 2015, ‘Sentiment analysis: Measuring opinions’, *International Conference on Advanced Computing Technologies, and Applications*, vol. 45, pp. 808-814.

10. M. Tubishat, N. Idris, M. A.M. Abushariah, 2018, ‘Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges’, *Inf. Process. Manage*, vol. 54, pp. 545-563.