Project Report

Over the last two decades innovations in data analytics, most notably neural networks, reinvigorated previously stale fields like natural language processing (NLP). Coupled with unprecedented amounts of new data (thank you internet), topics like sentiment analysis, document classification and question answering are off to the races.

However, even with almost limitless amounts of data and state-of-the-art techniques, NLP is not perfect. Text is often taken very literal. Irony and sarcasm are, for the most part, lost on neural networks. This is unfortunate as they often transform the meaning of a whole sentence. Consider for example “That’s great”. On its own, it sounds positive. Now let’s put it in context: “Oh, you weren’t paying attention at all. That’s great”. Suddenly the sentence has a negative connotation.

Literature review

Detecting and understanding sarcasm is one of the last hurdles in language processing. Research has mostly focused on semi-supervised learning approaches based on Twitter posts. Twitter lets users annotated their submissions with different hashtags (describing the content), one of which is “#sarcasm”. This makes it an excellent data source. Ptáček et al. (2014) demonstrate the viability of this approach achieving good results using Support Vector Machines.

Calzolari (2014) builds on this idea by using the “#sarcasm” tag to identify other tags that indicate sarcasm. Regarding sentiment analysis, he notes that just detecting sarcasm is not enough. Instead, the scope must be considered as well. Is the meaning of the whole sentence reversed or just the last few words? He closes by pointing out that the “#sarcasm” tag is unreliable.

It is added by the user, most of which do not know it exists while others use it in the wrong context. Labeling posts by hand can improve the results (Davidov et al. 2010) but is also very time and cost intensive. A different approach for data generation is needed.

For years Twitter has allowed multimodal submissions such as images. True to the saying “a picture is worth a thousand words” these images often capture the true meaning of a post better than the accompanying text. The sentence may be sarcastic, but the picture still shows the intended sentiment. A few researchers have caught on to this idea and incorporated images into their sarcasm detection (Schifanella et al. 2016; Das and Clark 2018). However, they still rely on hashtags to identify sarcastic tweets.

In contrast, we propose to use semi-supervised learning based on traditional sentiment analysis to identify tweets where the image reflects a different sentiment than the text, thus indicating sarcasm.

Methodology

By labeling images with the sentiment score of the accompanying text, most images will be annotated correctly. Only in case of sarcasm will the label be wrong. However, because of the relatively small number of sarcastic posts, this will have little impact. Once the model learned to assign a sentiment score to an image, sarcastic tweets will end up with opposing scores.

Data preparation

Two datasets are required to implement this idea. The first one contains text-based posts annotated with sentiment labels. Several similar datasets have been generated over the last few years and are freely available on the internet. We chose the “Sentiment140” set which contains 1.6 million tweets, half of which are positive while the other half is negative.

The second dataset must contain both images and text. Unfortunately, as of right now no such dataset is available online. However, the Twitter API can be leveraged to return only tweets containing pictures. Using a Docker container to listen for new posts yielded around 500,000 tweets per day. After filtering out non-English posts and inappropriate images, around 30,000 tweets remained.

Text in both datasets is separated into words. Every time a “not” is encountered the remaining words of the tweet are preceded with a “not_” (“not very good” => “not_very not_good”). Only words that occur in at least 50 tweets are kept. Images are resized to 227x227 and cropped if necessary.

Traditional sentiment analysis

The text is fed through an embedding layer followed by a 3-layer LSTM with 30% dropout. The encoded sentences are run through two linear layers with another 30% dropout in between. Lastly, the prediction is made by a SoftMax layer. The dropout was added after earlier iterations heavily overfitted. This architecture results in a 79% validation accuracy which is good enough as a proof-of-concept. A more complex model based on BERT (Devlin et al. 2018) was also built if more robust results are needed but could not be run as part of this project due to the hardware requirements.

The posts from the second dataset were prepared the same way and then a sentiment score was predicted for each Tweet. 7359 posts received a rating below 0.5 (negative), the remaining 25588 posts received a rating above 0.5 (positive). This class imbalance is unfortunate but not an issue for a proof-of-concept.

Image sentiment analysis

Users can post virtually anything on Twitter. This makes training on images very complicated. In typical classification tasks (i.e. ILSVRC) images of the same classes are similar in the sense that they depict the same object. For sentiment analysis, to completely different pictures can convey the same sentiment. Furthermore, the number of possible motives is infinite. Therefore, the classification algorithm has to focus on broader themes like tone and style.

To realize such an algorithm a custom architecture needs to be developed. Sadly, this task vastly exceeds the scope of this project. Instead, a pretrained AlexNet (Krizhevsky et al. 2012) was refitted using the 30,000 labeled images. The only addition to the network is a final Sigmoid layer to produce a sentiment score between zero and one.

Training on a GPU for 8 epochs resulted in an MSE-loss of 0.0515. Computing accuracy is difficult because the network will never return the exact sentiment score. Defining “correct” as any score that is at most 0.2 away from the label, the network achieved an accuracy of 61%. If the score is converted into binary labels (positive/negative), the accuracy increases to 78%.

Sarcasm detection

The validation set contains 6587 Tweets. Of these, 1501 (23%) end up with opposing sentiment scores when comparing the results from the two models. However, many of these are edge-cases where the score is around 0.5 and slight variances are enough to change the label. 212 Tweets (3%) have scores which differ more than 0.5. These are strong suspects for sarcasm.

Conclusion

The approach outlined in this work is a proof-of-concept for a semi-supervised learning algorithm based on Twitter posts containing images. The results are encouraging and warrant further research. The wide variety of images encountered in posts is an issue that can be solved by a specialized convolutional network. The overall accuracy can also be improved by using a more advanced sentiment classification network. However, this work presents a very good first step towards a new approach for sarcasm detection.

Publication bibliography

Calzolari, Nicoletta (2014): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, May 26-31, 2014. European Language Resources Association (ELRA). Available online at http://www.lrec-conf.org/proceedings/lrec2014/index.html.

Das, Dipto; Clark, Anthony J. (Eds.) (2018): Sarcasm Detection on Flickr Using a CNN: ACM.

Davidov, Dmitry; Tsur, Oren; Rappoport, Ari (Eds.) (2010): Semi-supervised recognition of sarcastic sentences in twitter and amazon: Association for Computational Linguistics.

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (2018): Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*.

Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E. (Eds.) (2012): Imagenet classification with deep convolutional neural networks.

Ptáček, Tomáš; Habernal, Ivan; Hong, Jun (Eds.) (2014): Sarcasm detection on czech and english twitter.

Schifanella, Rossano; Juan, Paloma de; Tetreault, Joel; Cao, Liangliang (Eds.) (2016): Detecting sarcasm in multimodal social platforms: ACM.