NLP Coursework
Predicting the funniness score of edited headlines

Cesare Magnetti **Antonio Vespoli** **Lisa Alazraki**

Abstract

We illustrate our approach to the regression subtask of Task 7 of the CodaLab SemEval-2020 competition "Assessing Humor in Edited News Headlines". Given a paired dataset of edited news headlines and their crowd-sourced humorousness ratings, we attempt to predict the funniness score of new headlines in the test set. First, we analyse the data thoroughly borrowing insights from cognitive science and linguistics. Next, we investigate the performances of several models with and without pre-trained embeddings. Lastly, we discuss the outcomes of all approaches and provide theoretical and practical insights behind their performance differences.

1 Code

The source code of the implementations discussed in this report is available at https://gitlab.com/cesare.magnetti/assessingheadlinesfunniness

2 Motivation

Humor and laughter are salient aspects of human communication, cognition and social interaction. It is therefore desirable for autonomous and conversational agents – as they become increasingly ubiquitous in society – to be able to understand humor and appropriately respond to it. This is a challenging task, as humor has been shown to be highly subjective and not easily quantifiable by humans themselves (1). This fact is especially true in the case of written language, where misunderstood humor often generates social-media drama, at times with significant consequences. It is therefore of great scientific interest to investigate automatic humor assessment and its wide range of applications: from avoiding misunderstandings in written communication to detecting sarcasm in NLP corpora that may lead to biased predictions in popular tasks such as Machine Translation (MT) and sentiment analysis.

3 Dataset

All the experiments have been performed on the Humicroedit dataset (2; 3). This corpus contains 15,095 news headlines collected from the Reddit forums `r/worldnews` and `r/politics`. The headlines have undergone minor edits with the aim to increase their humorousness; in most cases a single word was replaced, whilst a minority of samples received broader alterations. The dataset retains the original headline as well as the edits in a separate column. Moreover, each edited headline in the corpus is associated to funniness scores assigned by human judges – integer values between 0 ('not funny') and 3 ('funny') – as well as the real-valued average of these scores.
The vocabulary size of the tokenized corpus is 12138 unique tokens, reduced to 11820 after removing punctuation and stopwords.

4 Data analysis and exploration

A preliminary analysis of the dataset reveals a skewed grade distribution. The grades assigned to the samples in the train set have a mean of 0.94, a median of 0.8 and a standard deviation of 0.58. In other words, the majority of the samples are associated to low funniness scores, with few headlines being awarded a mean grade above the 1.5 range. Further investigation confirms that the mean grades in the test set follow the same distribution.

A significant portion of the dataset was collected from the subreddit `r/politics`, which is predominantly focused on U.S. internal affairs. We noticed that over a third of the headlines in the training set describes events relating to Donald Trump. Being the former American president a somewhat controversial and unpredictable figure, and since research into human neural responses and cognition associates unexpectedness with laughter (4), we hypothesised that headlines pertaining to Trump could have gained higher-than-average funniness scores. We isolated from the train set those examples where the word 'Trump' appears, either in the edit or in the original context, and we found that this subsample of data has a slightly greater average grade (mean 1.02, median 1.0) compared to the entire train set. Albeit small, the difference in the average funni-

ness score is not entirely negligible if we consider that the majority of the samples in the dataset have meanGrade values between 0.2 and 1.4 approximately.

Another conjecture stemming from the relation between incongruity and humor that has been observed in scientific literature is that a headline may be funnier when it contains at least one unexpected, out-of-context word (5). We tested this hypothesis by computing the vector norm on the GloVe embeddings (6) between the replaced word and the newly inserted one. This difference should be a measure of how unexpected the edit is in the original context. Therefore, we evaluated the correlation between this metric and the mean grades of the samples. Indeed we found a trend indicating that smaller distances between the original and the edited word correspond to lower meanGrade scores. In other words, replacing a word with another that is semantically similar is unlikely to produce funny headlines. The box-and-whisker chart below shows that, when the distance between the two words is at the lower end of the interval (between 1 and 3), it is unlikely for a headline to obtain a score higher than 1.2 approximately. On the other hand, as the distance between the removed word and the added one increments, both the mean score (represented by a dotted green line in the chart) and the median score (represented by a solid green line) increase, and the whiskers extend upward, indicating that there are samples within those distance ranges that obtained the highest funniness scores. These results seem to confirm our intuition that an edit too coherent with the sentence is less likely to result in a funny headline, whereas an edit that adds an element of surprise and incongruity has a higher chance of producing a comical effect.

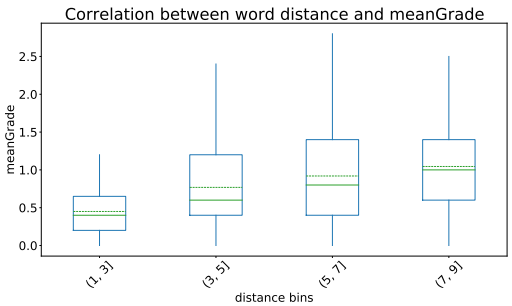


Figure 1: Correlation between the distance of the edit from the original word and the meanGrade score obtained.

Next, we briefly investigated the relationship between humorousness and cultural taboos (5). Using WordNet (7), we created a list of 262 taboo words, pertaining to human sexuality and physiology, by finding synonim sets (synsets) of a pre-determined array of starting concepts and manually removing from the final list those terms whose meaning deviated too significantly from the intended domain. We then identified 330 training samples where the edit column contained a taboo word.

We found that the headlines with taboo edits present, on average, a higher funniness score (mean 1.09) compared to those where the edit is non-taboo.

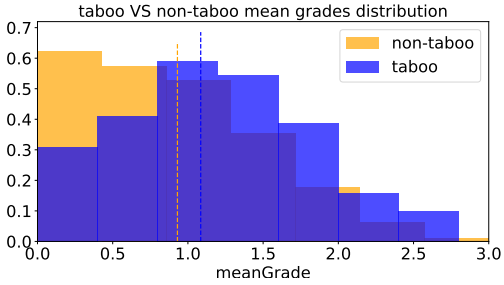


Figure 2: Comparison of the grade distrbutions for the taboo and non-taboo subsets in the data.

Although the limited amount of headlines containing taboo edits does not allow us to draw statistical comparisons between the two subsets, we can get the intuition that the headlines involving taboo words may be on average funnier than the others. This also connects back to the fact that taboo words are usually unexpected in the context of news headlines.

Lastly, we also analysed the variance between the grades given by each judge. We found that often judges did not vote unanimously, with the mean standard deviation being 0.72. This confirms that the concept of funniness is intrinsically subjective.

Ultimately, aside from the positive correlation found between the meanGrade score and the distance between the original and the edited word, the other experiments described in this section did not produce significant enough results to be integrated with the statistical and neural methods illustrated in the next paragraphs. Nevertheless, they were useful to understand the grade distribution and gain precious insights about the corpus.

5 Methods

In this section, we discuss two different approaches to the given task. In Approach 1, we investigate Statistical Machine Learning (ML) methods as well as Deep Learning (DL) models using pre-trained GloVe word embeddings (6), while in Approach 2 we train our own embeddings in an end-to-end fashion, and propose methodologies that fully avoid the use of word embeddings. All relevant results are reported and discussed in Section 6. Details regarding the architectures of each trained model can be found in the linked repository.

5.1 Baseline

We define our baseline as the method that assigns to each data point the overall average of all the meanGrades in the corpus, regardless of how funny a headline is. Interestingly, this method achieves reasonable results due to the fact that the distribution of the data is narrow around its mean.

5.2 Approach 1

Distributed word embeddings such as GloVe revolutionised the field of natural language processing, as they allow to capture the contextual semantics of words in a low-dimensional latent space. This key property enables the use of machine learning models to solve a variety of linguistic problems. We first considered more classical approaches, such as Random Forests (1), Support Vector Machines (8), and others. For each model we performed a 5-fold cross validation over a grid search of possible configurations. The input to such models should be a vectorized representation of the headline; this is achieved by computing the average of the embeddings of the words in the sentence. Note that this method gives equal importance to each token in a piece of text and, most importantly, it completely disregards the sequential nature of the input sentence. On the other hand, neural methods such as Recurrent Neural Networks (RNNs) (9; 10) and Transformers (11) are able to more rigorously capture the relations between different parts of a text. Given the high number of hyper-parameters of those models, a grid search would have been too computationally expensive. For this reason, we have selected the best model configurations using educated guesses. Another important consideration is that those models have a substantially higher number of parameters compared to statistical models and are more likely to over-fit to the training data. Given the insights gained in Section 4 suggesting funniness may be related to how much out of context an edit is, we experimented by adopting two different forms of input: (1) the edited tokenized headlines only, and (2) the concatenation of the original and the edited headlines, allowing the models to figure out such relationships. Finally, we compared results between vanilla input headlines and pre-processed input headlines. While in the former case we left the headlines unaltered, in the latter case we removed stop-words and punctuation. It must be noted that this step may influence the perceived funniness of a sentence.

5.3 Approach 2

Although it is usually preferred to adopt pre-trained word embeddings in data scarcity conditions, the aim of this section is to explore methods that do not take advantage of any form of pre-training. We tackled this complication of the problem with two different methods: (1) we added a randomly initialized embedding layer to the neural models, and (2) we adopted a strategy that does not rely on the use of word embeddings at all.

In the first method, the objective is to learn the word embeddings from scratch. Note that this modification significantly increases the number of tunable parameters and, therefore, we expect the training corpus to be too small for such models to learn meaningful representations.

In the second method, the concept of word embedding is completely avoided. Here, the headlines in the test set were compared to the headlines in the training set and a score was assigned based on cosine similarity. Those scores were obtained by converting the test headlines into vectorial representations called Bag of Words, and by doing the same for the headlines in the training set. Next, the normalized similarity between the unweighted test representation and the representations in the training set was calculated. The row in the training set with the highest score was chosen and the aligned mean grade was produced as output. A random choice was made if there were several training headlines with the highest score. As a second step, we evolved this approach by selecting the K closest instances and computing the average of the associated mean grades. In a further exploratory step, we experimented with some of the evaluation metrics typically employed in Machine Translation (MT) by using them as measures of distances for selecting the K closest instances. We adopted BLEU (12) and METEOR (13) as they correlate highly with human evaluation. The rationale for preferring these metrics is that they can better detect similarities at semantic level compared to cosine similarity. We found this method to be less effective than those relying on the use of word embeddings. This technique produced worse results compared to the baseline and converged to the baseline RMSE increasing K. However, since this method relies on a retrieval-based approach, it could become significantly more effective in the case of sizable corpora.

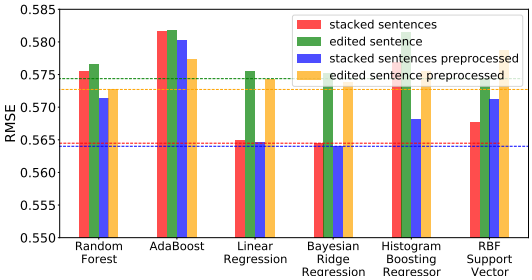
Another set of experiments was performed by creating a bag of words representation for each headline and training statistical models on them. We found that the best performing model is a linear regression with bag of words of 300 dimensions.

6 Results

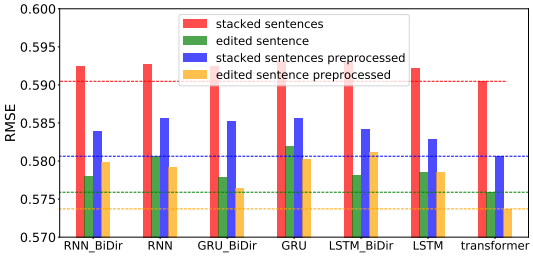
We experimented with emdeddings of different dimensions (50d, 100d, 200d and 300d), and found that the best performing size is 50d. Figure 3a reports the RMSE of the respective statistical models, while Figure 3b shows the same plot for the respective neural approaches. We found that the best performing statistical model is a Bayesian Ridge Regression trained on stacked edited and original headlines (RMSE ≈ 0.564). The top performing neural model is a transformer encoder trained on the edited headlines only (RMSE ≈ 0.573). In both cases results were better when stopwords and punctuation had been removed.

Training our own word embeddings improved the performance of the neural models in some cases and worsened it in others. This may be a symptom of unstable training due to the large number of parameters. In Figure 3c, we show the RMSE associated to the respective neural models without using pre-trained embeddings. The top performing model was a transformer trained on the edited headlines only, with punctuation and stopwords removed (RMSE ≈ 0.0558). An interesting ob-

a) Statistical ML models results



b) DL models results



c) DL models results, no pre-trained embeddings

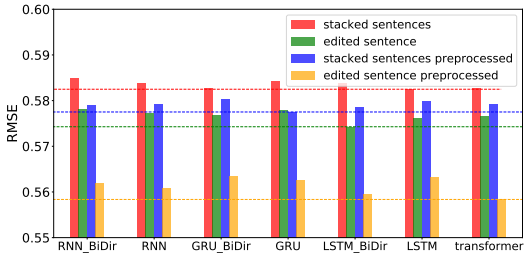


Figure 3: Achieved cross-validated RMSE errors for all models trained in Approaches 1 and 2.

servation is that those models were found to be highly dependent on their random initialisation. In Table 1, we collected the results on the test set of the best performing models for each category.

	RMSE	R2
Baseline	0.578	0.
Bayesian Ridge + GloVe	0.558	0.071
Transformer	0.579	0.
Transformer + GloVe	0.578	0.005
BoW + Linear Regression	0.578	0.008

Table 1: Results of the best models on the test set.

Overall, The Bayesian Ridge regression was found to give the lowest RMSE and the highest R2 score. It is important to note that all R2 scores are significantly low, which means that, although the models achieve low RMSE, the predictions distribution does not match the actual data distribution. Being the data extremely skewed, all models tend to produce outputs between 0.7 and 1.2. This is because the models are indirectly learning to infer funniness scores close to the mean score of the train set. Figure 4 visualises this issue for the Bayesian Ridge. Moreover, the predictions produced by the neural models are strongly clustered around the mean. This observations suggest that, despite the low RMSE scores, none of the models presented above properly captures the relationship between a headline and its funniness.

Output distribution of Bayesian Ridge model

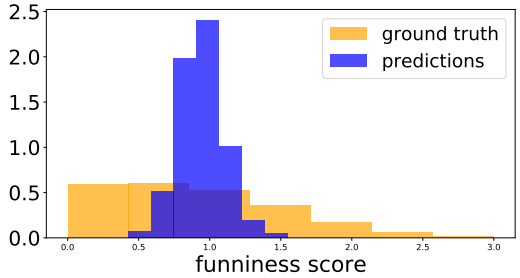


Figure 4: Predictive distribution of the Bayesian Ridge regressor, indicating that the model tends to cluster predictions around the mean grade of the train set (≈ 0.94). This issue is reflected by the low R2 score.

7 Conclusions

In conclusion, we explored with many different approaches and technologies the problem of predicting a funniness score given an edited headline. From our analysis of the corpus, we found that the distribution of the mean grades is significantly skewed. In particular, we observed that this leads both statistical and neural approaches to fail in predicting the meanGrade for the most funny headlines.

In terms of future developments, an interesting further investigation could be the use of transfer learning techniques adopting pre-trained language models such as Sentence-BERT (14). Applying it in the encoder would lead to headline representations that better capture their semantics. Another observation is that jokes may be related to specific events or recent news. For instance, many jokes about Donald Trump assume that the reader has some knowledge about him and his character. In order to deal with this type of situations, we could exploit techniques for integrating contextual information by defining a context vector built with information from other sources (e.g. recent news, biographies, etc). Moreover, some data augmentation techniques, such as replacing words with synonyms or adding some noise to the mean grades, could help make the models more robust and may be a way to compensate for the significantly unbalanced nature of this corpus. Some possible techniques for finding synonyms include using WordNet, finding the closest embeddings through GloVe, or performing back translation (i.e. using a MT system for translating into another language and back to English).

References

[1] Raskin V. Semantic mechanisms of humor. vol. 24. Springer Science & Business Media; 2012.

[2] Hossain N, Krumm J, Gamon M. "President Vows to Cut <Taxes> Hair": Dataset and Analysis of Creative Text Editing for Humorous Headlines. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 133–142.

[3] Hossain N, Krumm J, Gamon M, Kautz H. SemEval-2020 Task 7: Assessing Humor in Edited News Headlines. Arxiv preprint. 2020.

[4] Vrticka P, Black JM, Reiss AL. The neural basis of humour processing. Nature Reviews Neuroscience. 2013;52(1):29–65.

[5] Brône G, Feyaerts K. Introduction: Cognitive linguistic approaches to humor. Humor - International Journal of Humor Research. 2006;19(3):203–228.

[6] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532–1543.

[7] Fellbaum C, editor. WordNet: an electronic lexical database. MIT Press; 1998.

[8] Bishop CM. Pattern recognition and machine learning. springer; 2006.

[9] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–1780.

[10] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078. 2014.

[11] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv preprint arXiv:170603762. 2017.

[12] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a Method for Automatic Evaluation of Machine Translation. 2002 10.

[13] Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, Michigan: Association for Computational Linguistics; 2005. p. 65–72.

[14] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks; 2019. p. 3973–3983.