SemEval-2017 Task 6: Humor Detection

Paul Craig
craig263@d.umn.edu

1 Introduction

Sem-Eval 2017 Task 6, tasks it's participants with creating a program capable of classify tweets as humorous by the standards of the show Hashtag wars. The show in question has a segment where each week the top 10 funnest tweets for a given hashtag are presented. Participants where given a set of annotated tweet sets from the show. Each tweet set contains tweets from the shows top ten as well as a number of none placing tweets. The tweets are annotated with a number showing their relative placement by the show. The tweet for a hashtag ranked number one by the show is annotated with a score of 2. The other top nine are annotated with a score of 1. Finally tweets that where not ranked in the top ten are scored with a 0. All in all there are 12734 tweets in the task. 11325 of these are reserved as training data, 749 are for evaluation, and 660 being used as the trial data for the task(Potash et al., 2017).

Sub-task A of the Sem-Eval gave participants two tweets from the show. Participants had to create a program capable of determining which of the two tweets was funnier. For the the purposes of the task the funnier tweet is the one which received a high score. For trial data all tweet pairs where guarantied to have different scores. Given the binary nature of this task a baseline score 0f 50% accuracy is expected given a random classifier.

2 Review of Other Teams Approaches

2.1 Humor Hawk

David Donahue, Alexey Romanov, and Anna Rumshisky of the University of Massachusetts Lowell competed in the Semeval under the team Name Humor Hawk. They placed first in Sub task A with an accuracy of .675 on their second attempt. The team solved the problem using a combination of several models. The first is what the team Refers to as a Character-to-Phoneme Model. This model translates words in to a sequences of Phonemes that represent the word based of how it is pronounced rather than how it is spelt. The output of this model is then passed to an embedding layer which encodes words using pre-trained Glove vectors and a LSTM neural network. The team also created an embedding model based of the character representation of the tweets. The outputs of the Phonemes embedding and the character embedding where then fed into dense layers and combined using a convoluted neural network in to an Embedding/Character Joint Model. In addition to their machine learning models the team also created a rule based model that took into account several factors including sentiment analysis, tweet length, and POS. The rule based system and the output the Joint Model where combined into a finale ensemble model which produced the final prediction(Donahue et al., 2017).

2.2 DataStories

A team from the University of Piraeus competed in the SemEval under the team name DataStories. The team had the third best accuracy overall for sub-task A.

The first step in the teams method is to generate word embeddings using Glove on a twitter data set(Pennington et al., 2014). Next the team prepossessed the input text with a custom tokenizer. This tokenizer spell checks and segments the tweets, and also handle emoticons. After preprocessing text is passed to a Siamese LSTM neural network. The neural net is subdivided into two sub-networks with shared weights. Each of theses sub-networks is then given one of the two tweets for comparison. Each sub-net has three layers. First is an Embedding Layer which projects tweets into vector space using the team Glove embedding. Next is a BiLSTM layer which transitions

between hidden states. Finally data goes though a context Attention Layer which assigns weights to words based on it perceived importance. The output of both sub-networks are then feed into a Fully-Connected Layer which combines both outputs into a single representation. This representation is then feed to a final output layer which performs a binary classification to decide which tweet is funnier(Baziotis et al., 2017).

2.3 Duluth

Xinru Yan and Ted Pedersen of the University of Minnesota Duluth competed in the Sem-Eval under the team name Duluth. They ranked fourth overall in terms of accuracy. This team method revolved around using language models to determine how similar tweets where to other tweets, and how dissimilar they where to news articles.

The team first trained a pair of Language models. The first was trained off of a corpus of tweets. The second was then trained off a corpus of news articles. Each LM was trained using the KenLM Toolkit. These LM where then use to classify the trail data

Before being used by the LM's trial data went though prepossessing. First unnecessary characters such as '@' or Urls where striped form the data.Tweets where then separated out into their individual tokens by splitting on white space and punctuation.

Tweets where then assigned a log probability based on there similarity assigned by each language model. Tweets where considered to be funny if they where assigned a high probability by the tweet based language model and a lower probability based from the news language model(Yan and Pedersen, 2017).

3 Purposed Method

For my attempt of the sem-eval I used a the distilbert model provided by hugging face, along side some minimal pre-processing.

3.1 Pre-Processing

In pre-processing the hashtag token in each tweets was extracted and replaced with a space separated version. The hashtag was then distinguished from the rest of the tweet by surrounding it with a starting ¡hashtag¿ and an ending¡/hashtag¿ token. for example the hashtag 420Celebs would become: ¡hashTag¿ 420 Celebs ¡/hashTag¿. The tweets for each hashtag where subdivided by their scores. Each tweet for a hashtag was paired with every other tweet in the hashtag with a different score. Two version of these pairs where then saved as input examples for the network. The first example puts the higher scoring tweet first and is given a label of 1. The second example put the lower scoring tweet first are given a label of 0. These examples where then used to fine tune the hugging face model for subTask A.

3.2 Classification

The classification task was carried out by a pre-trained distilbert model provided by the hugging face transformers package (Wolf et al., 2019). I chose to use the distlibert model as it smaller number of parameters (as compared to bert) allowed for more test to ascertain the effectiveness of various pre-processing strategies and meta parameters. The finale model was trained over the course of 3 epochs with a batch size of 32, and an initially learning rate of 2e-5.

4 Results

The final system achieved an accuracy of 61% as calculated by the official evaluation script released with the task. This put it below the result of the other teams covered in this paper (Humor Hawk: .675, Datastories: .632, Duluth 6.27) but still well above the 50% baseline. I suspect that this disparity in scores is due to the lack of features such as POS tags and normalization of emoticons and urls. These features where present in both the Duluth and Data stories systems, but had to be cut from mine due to problems with the tokenizer.

5 Ethical consideration

In developing a system for detecting humor there are several ethical considerations that must be taken. The first is the unfortunate fact that humor has often been used as a shield for unacceptable behavior. The defense of hate-speech being "just a joke" has history of being used to justify and normalize hate speech (Billig, 2001) As such we must consider the possibility that a humor classifying program could give unintentional validation to these acts if it labels such hate speech as humor. Out side of validating hate speech, a false positive in humor could prove to be problematic if applied to safety information. A real life example of this can be found in the Senton Hall dorm fires. Senton

Hall had a history of their fire alarm being pulled as a prank. Thus when there was an actual fire many of the students assumed it was a joke and went back to sleep(Steinberg, 200AD).

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 task 6: Siamese LSTM with attention for humorous text comparison. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 390–395, Vancouver, Canada. Association for Computational Linguistics.

Michael Billig. 2001. Humour and hatred: the racist jokes of the ku klux klan. *Discourse Society*, 12.

David Donahue, Alexey Romanov, and Anna Rumshisky. 2017. HumorHawk at SemEval-2017 task 6: Mixing meaning and sound for humor recognition. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 98–102, Vancouver, Canada. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57, Vancouver, Canada. Association for Computational Linguistics.

Martin Steinberg. 200AD. 3 die in seton hall dorm fire. *Seattle Times*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Xinru Yan and Ted Pedersen. 2017. Duluth at semeval-2017 task 6: Language models in humor detection. *CoRR*, abs/1704.08390.