#WarTeam at SemEval-2017 Task 6: Using Neural Networks for Discovering Humorous Tweets

Iuliana Alexandra Fleșcan-Lovin-Arseni, Ramona Andreea Turcu, Cristina Sîrbu, Larisa Alexa, Sandra Maria Amarandei, Nichita Herciu, Constantin Scutaru, Diana Trandabăț, Adrian Iftene

University Alexandru Ioan Cuza of Iași, Romania
{alexandra.flescan, ramona.turcu, cristina.sirbu, larisa.alexa,
sandra.amarandei, nichita.herciu, dtrandabat, adiftene}@info.uaic.ro

Abstract

This paper presents the participation of #WarTeam in Task 6 of SemEval2017 with a system classifying humor by comparing and ranking tweets. The training data consists of annotated tweets from the *@midnight* TV show. #WarTeam's system uses a neural network (TensorFlow) having inputs from a Naïve Bayes humor classifier and a sentiment analyzer.

1 Introduction

One of the most recent direction in Artificial Intelligence is related to humor and, in recent years, comedy based computing such as Manatee (Gustin, 2014), the joke writing computer, STANDUP - System to Augment Non-Speakers' Dialogue Using Puns (Waller et al., 2009); SASI the sarcasm-detector (Davidov et al., 2010), or DeviaNT (Kiddon and Brun, 2011) were developed, with more or less success (Leybovich 2017). If the well-hidden structure of humor, from which are derived all uncertainties, would be uncovered, it would have great applicability in social networks and human computer interactive systems. In time, research has been made and progress is undeniable. However, most recent studies are concerned with a binary perspective over humor where two main features are ignored: its continuous nature and subjectivity.

Our objectives in Task 6 of SemEval 2017 (Potash et al., 2017) were: (1) to build an application able to score the degree of humor in tweets from the Midnight TV show, the Hashtag War section and (2) to discover ways to automatically determine amusement and how to quantify it.

The paper is structured in 5 sections: Section 2 discusses existing approaches to humor detection and Section 3 presents the methodology of our system. Section 4 briefly analyses the obtained results, before Section 5 drafting some conclusions and further work.

2 State of the Art

In the area of identifying, describing and evaluating humor, the majority of studies succeeded only to describe if something is funny or not. The actual tendency is to move forward to something more specific, namely to the value or the degree of humor. Currently, studies are mainly concerned with the binary evaluation of humor, whether it is funny or not. Their object of study is different as some of them focused on evaluating humor in videos and images, while others in texts expressed in natural language.

As for the studies related to identifying humor in pictures (Chandrasekaran et al., 2016), theories in this area suggest that humor's key components are qualities such as unexpectedness, incongruity, pain, as observed by analyzing a database of 6,400 funny and not funny images.

The linguistic side of this computational approach identifies the mechanisms for humor detection with a formal model of the semantic and syntactic regularities underlying some of the simpler types of punning riddles (Mulder and Nijholt, 2002).

Barbieri and Saggion (Barbieri and Saggion, 2014) represents the task as a classification problem, applying supervised machine learning methods taking into account a group of features: frequency, written-spoken style uses; intensity of ad-

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 407–410,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

verbs and adjectives; structure (length, punctuation, emoticons, links), sentiments (gap between positive and negative terms); common vs. rare synonyms use; ambiguity (measure of possible ambiguities). In a part of their research, they treat irony and humor as a single class called figurative language and, by using specially designed humor characteristics, obtain accuracy around 76%.

A similar direction is investigated in (Yang et al., 2015), where they first formulate the task as a traditional text classification problem, to further apply Random Forest. At the same time, semantic structures behind humor are analyzed in terms of meaning incongruity, ambiguity, phonetic style and personal affect. A simple and effective method of Maximal Decrement is proposed. The phonetic style (alliteration, rhyme, word repetition etc.) of a joke is regarded as being at least as important as its content.

Several studies agree that humor has at its basis incongruity. In (Mihalcea et al., 2010), models are analyzed based on their features:

(1) semantic relatedness, where the intuition is that the correct punch line will have a minimum relatedness with respect to the set-up: knowledge-based metrics and corpus-based metrics (vector space model and pointwise mutual information), based on word co-occurrence over very large corpora, and domain fitness obtained from WordNet domains; and

(2) joke-specific features: polysemy and latent semantic analysis trained on joke data that contains one-linears (short sentences with comic effects, simple syntax, rhetoric devices and creative language constructions). As the authors confess, the difficulty in detecting incongruity is that it has to satisfy to opposite requirements, namely to be coherent but to produce a surprising effect. Using a combined model consisting of an SVM learning system trained on a combination of knowledge-based, corpus-based, and joke-specific features, they obtained a precision of 84%.

The most common and efficiently used text classifiers are Naive Bayes and Support Vector Machines (Mihalcera and Pulman, 2007).

The first one is used to estimate the probability of a category using joint probabilities. The second ones are binary classifiers that seek to find the hyperplane that best separates a set of positive examples from a set of negative examples, with maximum margin.

Our approach proposes the use of neural networks as an interface between a Naïve Bayes classifier and a sentiment analyzer, trained on the data provided by SemEval 2017 task 6 organizers.

3 Methodology

#WildDev's team developed a system for classifying humor by comparing and ranking a set of tweets on the basis of a collection of hashtags from @midnight show. Our approach considered using two machine learning techniques: neural networks and Naïve Bayes.

The format of the trial and training data was established by the task organizers (Potash et al., 2017), having the following structure:

“720293211374104578 Honey, I lost the house. #VegasMovies @midnight 0”

with a tweet ID, the text of the tweet, the hashtag it related to in the #HashtagWar at @midnight show, and a score.

Each tweet in a set of tweets is evaluated with a score of 0, 1, or 2, where 2 corresponds to the funniest tweet in the set, 1 corresponds to a tweet in the top 10 funniest tweets, and 0 corresponds to a tweet not in the top 10 funniest tweets (most of the tweets in a file). This way, the continuous nature of humor can be investigated.

The architecture of the #WarTeam is presented in figure 1 and includes four modules: a pre-processing task; a Naïve Bayes classification algorithm for identifying humorous vs. non-humorous instances; a simple, dictionary-based sentiment analyzer and a neural network supervised algorithm. Each specific module is further detailed below.

3.1 Pre-processing

The first module consists in a pre-processing phase, a component responsible with cleaning each tweet before passing it to the machine learning algorithms. The goal of this module is to remove unneeded data which might have a bad impact on the learning algorithm.

The pre-processor module is a JAVA standalone application that receives as input a file with multiple tweets, one per line, and returns a list with processed tweets. Several rules are applied in the process of cleaning tweets. The most important one is removing frequent hashtags.

We consider a hashtag to be frequent when it appears in at least two of the tweets given as input. This rule was established due to our belief

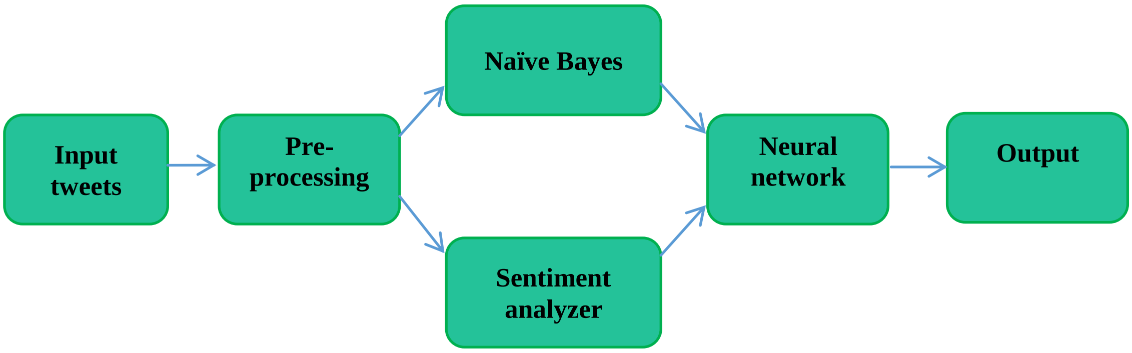

Figure 1. System Architecture

that humor, in the sense of the TV show *@midnight* from Comedy Central, from which the training data was collected, arises mostly from new, creative content.

If the hashtag is unique, the pre-processor will try to split it into separate words (this rule applies only for hashtags written with camel case).

After all hashtags have been processed and any other irrelevant data removed (e.g links, punctuation), the tweets are lemmatized and each word is replaced with the corresponding lemma. This will help the learning algorithm find more matches in the list of tweets, than it would if word forms were used.

This module uses two language processing libraries: Stanford parser (Klein and Manning, 2003) for tokenization and WordNet[1] for finding a word's lemma.

3.2 Naïve Bayes

Two machine learning algorithms were used to extract humor from tweets: a Naïve Bayes and a neural network, trained on the data provided by SemEval 2017 Task 6 organizers.

The first solution we adopted was to train binary Naïve Bayes algorithms on the training data, for each category of scores. As features for the Naïve Bayes classifiers, we used data from the pre-processing module (lemmas)

However, the classifiers turned out to be rather biased, since only one most funny tweet (score 2) and top ten funniest tweets (score 1) are annotated for each set of tweets, and the majority of tweets have the score 0.

3.3 Sentiment analyzer

Trying to improve the results of the classifier, we developed a further module responsible for attaching a polarity score to each word in the tweet:

"<word>" : <polarity_score>,

This simple sentiment analyzer used a manually acquired dictionary of about 2500 lemmas annotated with a sentiment score ranging from -5 (corresponding to the extreme negative sentiment) to +5 (the extreme positive one). The words not included in this list were considered neutral and received the polarity_score 0.

Using a python program, the input tweets and the dictionary of polarity scores, a list of word pairs with corresponding scores was generated.

"<word1> <word2>" : <score>,

The main idea behind this approach is that there are contrastive bigrams more frequently indicating humor, such as "black milk".

3.4 Neural network

The output of the Naïve Bayes classifier, along with the scores generated by the sentiment analyzer, are inputs for a neural network algorithm. Additionally, a manually generated corpus of celebrity names was also used as input.

"<name>" : <score>,

This was motivated b the observation that tweets containing celebrity names were considered more attractive.

A neural network with 101 neurons was trained to rate the tweets in their final form. This algorithm can be used for a file or only for one tweet. Thus, each tweet will have a score from the different machine learning algorithms that represents a value on the 'funny' scale (greater value = funnier), in the [0,1] interval. Based on these scores, the tweets are ordered and the first tweet is awarded the final score 2, the next 9 tweets receive 1, and the rest a score of 0.

[1] Java WordNet Library from https://sourceforge.net/projects/jwordnet/

4 Evaluation

The running time for the pre-processing phase is less than 2.5 seconds and for the neural network and rating algorithms is less than 7.0 seconds per file for 101 neurons. The size of the created corpuses are: polarity scores dictionary with about 2500 words, bigram lists with about 2000 word pairs, and the celebrities corpus with 50 names.

The implemented neural network algorithm returns a label for every tweet provided as input. Testing the algorithm using training data provided (10 fold cross validation), the accuracy of the algorithm proved to be 10286 of 11325 tweets correctly identified (mostly the ones with a 0 score).

5 Conclusions

Classifying and ranking humor is certainly a challenging task. The major challenge comes from the subjective nature of humor and the influence of the cultural background on identifying humorous situations.

#Warteam participated in SemEval 2017 Task 6 with a system combining Naïve Bayes and neural networks. This participation was an excellent way to consolidate natural language skills, while being involved in an international competition.

For a first try, the results are satisfying, given the fact that our algorithm succeeded to identify humor rules similar to the ones identified by human while looking through training data.

Although our system needs improvements, the research interest for this field was open and progress was done. Taking this into consideration, the improvements we consider for our system implies: better scoring algorithm to provide higher credibility to either of the two learning algorithms we used for each individual file, not at a whole as we currently do; improve running time for the neural network, enrich the corpora using assisted automatic web crawling techniques, but also use an API to identify positive and negative sentiments.

But the most challenging research direction yet to be investigated is how to incorporate cultural background when classifying tweets for their humor.

References

Barbieri, Francesco, and Horacio Saggion (2014) *Automatic detection of irony and humour in twitter,* in Proceedings of the Int. Conf. on Computational Creativity.

Chandrasekaran Arjun, Ashwin K. Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh (2016) *We Are Humor Beings: Understanding and Predicting Visual Humor*, arXiv:1512.04407

Dan Klein and Christopher D. Manning. (2003). *Accurate Unlexicalized Parsing*. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Davidov, Dmitry, Oren Tsur, and Ari Rappoport. (2010) *Semi-supervised recognition of sarcastic sentences in twitter and amazon,* in Proceedings of the fourteenth conference on computational natural language learning. ACL, 2010.

Ilya Leybovich (2017) *Is Humor the Final Barrier for Artificial Intelligence?* from https://iq.intel.com/is-humor-the-final-barrier-for-artificial-intelligence/.

Kiddon, C., & Brun, Y. (2011). *That's what she said: double entendre identification*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 89-94). Association for Computational Linguistics.

M.P. Mulder, A. Nijholt. (2002) *Humour Research: State of the Art,* Technical Report CTIT-02-34.

Mihalcea, Rada, and Stephen Pulman. (2007) *Characterizing humour: An exploration of features in humorous texts* in Int. Conf. on Intelligent Text Processing and Computational Linguistics.

Mihalcea, Rada, Carlo Strapparava, and Stephen Pulman. (2010) *Computational models for incongruity detection in humour.* in Int. Conf. on Intelligent Text Processing and Computational Linguistics.

Potash Peter, Romanov Alexey and Rumshisky Anna (2017) *SemEval-2017 Task 6: #HashtagWars: Learning a Sense of Humor*, in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017).

Sam Gustin (2014) *It's Comedian vs. Computer in a Battle for Humor Supremacy*, from https://www.wired.com/2014/04/underwire_0401_funnycomputer/

Waller, A., Black, R., O'Mara, D.A., Pain, H., Ritchie, G., Manurung, R. (2009) *Evaluating the STANDUP Pun Generating Software with Children with Cerebral Palsy*. ACM Transactions on Accessible Computing (TACCESS) Volume 1, Issue 3 (February 2009) Article No. 16.

Yang, D., Lavie, A., Dyer, C., & Hovy, E. H. (2015). Humor Recognition and Humor Anchor Extraction. In EMNLP (pp. 2367-2376).