Recognizing Sarcasm in Twitter: A Comparison of Neural Network and Human Performance

David Kovaz
University of Memphis

Roger J. Kreuz
University of Memphis

Introduction

- A growing body of literature suggests that lexical (word-level) features may serve as important cues for distinguishing sarcasm from literal language (e.g., Kovaz et al., 2013).
- Lexical features have been successfully used in machine learning algorithms for sarcasm recognition in Twitter (Davidov et al., 2010; González-Ibáñez et al., 2011).
- Our goal is to build a neural network model based on lexical features that distinguishes sarcastic from nonsarcastic statements in Twitter.

Research Questions

- How well does the neural network perform in classifying sarcastic and nonsarcastic statements?
- How does this compare to human performance on the same task?

Corpus and Features

- Corpus consisted of pairs of tweets from 941 unique Twitter users collected by Kovaz et al. (2013).
- Each pair consisted of one tweet explicitly marked as sarcastic and one nonsarcastic tweet (see Figure 1):
  - Sarcastic tweet marked with #sarcasm.
  - Tweet from the same user not marked with #sarcasm was used as the nonsarcastic tweet.
- 101 lexical features extracted from each tweet (see Table):
  - 34 parts-of-speech tagged using the Stanford POS Tagger.
  - 65 categories from Linguistic Inquiry and Word Count (LIWC).
  - Presence of user references (@User) and hashtags (#hashtag).

Network Structure and Training

- Network consists of three layers (see Figure 2):
  - **Input layer:** Contains 101 nodes corresponding to the lexical features. Each feature coded as present (1) or absent (0).
  - **Hidden layer:** Distributed representation of the features consisting of 40 nodes. Activation values for each node range from 0 to 1.
  - **Output layer:** Contains a single node with an activation value ranging from 0 to 1. Returns a probabilistic judgment of sarcasm or nonsarcasm.
- Values for weights were randomized at the start of training.
- During training, the activation value of the output node was compared to a target value (1 for sarcastic and 0 for nonsarcastic) for each tweet run through the network.
- Weights were trained using backpropagation algorithm.

Figure 1: Example Pair of Sarcastic and Nonsarcastic Tweets

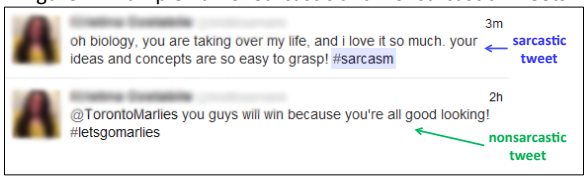
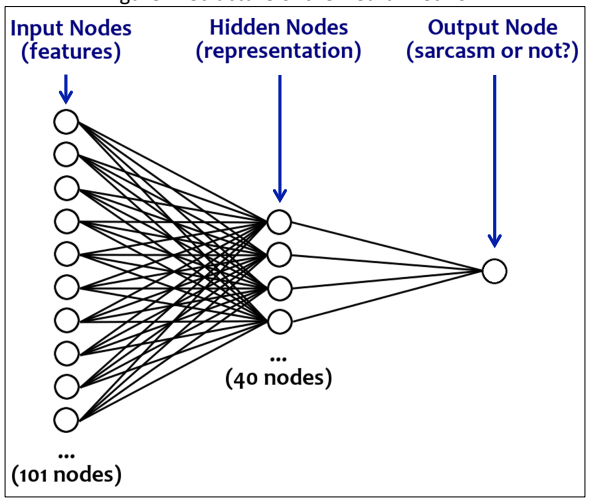


Table: Examples of Lexical Features

Part-of-Speech Tags (34)		LIWC Categories (65)	
Tag	Description	Category	Examples
DT	determiner	certain	always, never
JJ	adjective	family	daughter, husband
NN	singular noun	hedge	approximate, likely
PRP	personal pronoun	negemo	hurt, ugly
RB	adverb	posemo	love, nice
RP	particle	social	mate, talk
UH	interjection	swear	damn, piss
VB	verb	time	end, until

Figure 2: Structure of the Neural Network



Performance

Neural Network Performance

- Binary classifications (sarcasm vs. nonsarcasm) were made using an output node activation value cutoff of 0.5.
- We assessed the performance of the network using a 10-fold cross-validation procedure.
- The average accuracy of the network was 62%.

Human Performance

- We recruited 53 participants from Amazon Mechanical Turk to judge 100 sarcastic and 100 nonsarcastic tweets from the same corpus. They were given a list of tweets and asked to determine whether each tweet was sarcastic or not.
- The average accuracy of these human raters was 70%.

Network vs. Human Error Analysis

- We calculated the squared deviation of output from target value (error) for each tweet run through a trained network. We also calculated the average human rater accuracy for each tweet.
- Network error and human accuracy were significantly correlated ($r = -0.25$, $p < .001$) such that tweets with lower network error tended to be classified more accurately by human raters.

Discussion

- Our neural network model for classifying sarcasm from nonsarcasm in Twitter achieved an average accuracy of 62%. This was slightly below human performance (70%) on the same task.
- Error analysis suggests that the network and human raters made similar errors in classifying sarcastic and nonsarcastic tweets.
- These results underscore the importance of lexical features and show some promise for using neural network models in sarcasm recognition.
- Future models may use theory or data-driven feature sets and incorporate contextual information from streams of messages.

References

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. *Proceedings of the 14th Conference on Computational Natural Language Learning*, 107-116. Uppsala, Sweden.

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 581–586. Portland, OR.

Kovaz, D., Kreuz, R. J., & Riordan, M. A. (2013). Distinguishing sarcasm from literal language: Evidence from books and blogging. *Discourse Processes, 50*, 598-615.