

# Humor Detection in Product Question Answering Systems

Yftah Ziser  
Amazon  
Haifa, Israel  
yftahz@amazon.com

Elad Kravi  
Amazon  
Haifa, Israel  
ekravi@amazon.com

David Carmel  
Amazon  
Haifa, Israel  
dacarmel@amazon.com

## Abstract

Community question-answering (CQA) has been established as a prominent web service enabling users to post questions and get answers from the community. Product Question Answering (PQA) is a special CQA framework where questions are asked (and are answered) in the context of a specific product. Naturally, humorous questions are integral part of such platforms, especially as some products attract humor due to their unreasonable price, their peculiar functionality, or in cases that users emphasize their critical point-of-view through humor. Detecting humorous questions in such systems is important for sellers, to better understand user engagement with their products. It is also important to signal users about flippancy of humorous questions, and that answers for such questions should be taken with a grain of salt.

In this study we present a deep-learning framework for detecting humorous questions in PQA systems. Our framework utilizes two properties of the questions – Incongruity and Subjectivity, demonstrating their contribution for humor detection. We evaluate our framework over a real-world dataset, demonstrating an accuracy of 90.8%, up to 18.3% relative improvement over baseline methods. We then demonstrate the existence of product bias in PQA platforms, when some products attract more humorous questions than others. A classifier trained over unbiased data is outperformed by the biased classifier, however, it excels in the task of differentiating between humorous and non-humorous questions that are both related to the same product. To the best of our knowledge this work is the first to detect humor in PQA setting.

## CCS Concepts

• Information systems → Question answering.

## Keywords

humor detection, product question answering

## ACM Reference Format:

Yftah Ziser, Elad Kravi, and David Carmel. 2020. Humor Detection in Product Question Answering Systems. In *43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, July 25–30, 2020, Virtual Event, China. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3397271.3401077>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*SIGIR '20, July 25–30, 2020, Virtual Event, China*

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8016-4/20/07...\$15.00

<https://doi.org/10.1145/3397271.3401077>



Figure 1: A product titled "Think Geek Canned Unicorn Meat", associated with many humorous questions, e.g.: "Does it give you ever-lasting life?", or "Is it kosher?"

## 1 Introduction

Community question-answering (CQA) has been established as a prominent web service, enabling users to post questions and get answers from the community. eCommerce websites like Amazon<sup>1</sup> and eBay<sup>2</sup> maintain CQA platforms around products, supporting pre- and post-purchase inquiries. In these product question answering (PQA) platforms, questions and answers are posted in the context of a specific product, thus creating an ad-hoc community of users with common interest around the product. Naturally, a lot of effort is invested in managing content quality in PQA platforms, for example, some questions are boosted as they are relevant to many users, while others, e.g., offensive questions, are often removed.





Humorous questions lie in between; on one hand they engage users that are happy to enjoy some good laugh, but on the other hand, some users may find them confusing and even offending. For example, consider the canned unicorn meat box presented in Figure 1. One humorous question posted with respect to this product is "Is it kosher?". While some people may find it humorous, others may find it rude. Table 1 provides some more examples for products associated with humorous and non-humorous questions.

Humor detection has attracted a lot of research recently, especially with the emergence of deep-learning tools [3, 5, 6]. It has been shown [6, 20, 39] that associating text with context can improve humor classification accuracy significantly. While in some settings the context is unknown, which makes humor detection a challenge [20], PQA has essentially a built-in context – the product details. For example, the question "can it also be used for making coffee?" has a non-humorous intent when asked in the context of a tea pot, and a humorous intent in the context of a Swiss army knife. One of our goals in this work is to learn how the product context can be effectively utilized to improve humor detection in PQA systems.

<sup>1</sup>[www.amazon.com](http://www.amazon.com)

<sup>2</sup>[www.ebay.com](http://www.ebay.com)

**Table 1: Products associated with humorous and non-humorous questions**

Product	Image	Humorous Questions	Non-Humorous Questions
Nintendo Switch Gray Joy-Con		<ul style="list-style-type: none"> <li>• Can i use this to hack into the matrix and save humanity?</li> <li>• Can I trade one of my kidneys?</li> <li>• What if the princess wants to be with Bowser and Mario keeps kidnapping her?</li> </ul>	<ul style="list-style-type: none"> <li>• What do the ports on the side of the console do?</li> <li>• How much money will the system cost?</li> <li>• How do I know if this is the neon or gray version?</li> </ul>
Echo Show - 1st Generation Black		<ul style="list-style-type: none"> <li>• Will this help me find the meaning of life?</li> <li>• Can Alexa show me my future?</li> <li>• Does it cook breakfast?</li> </ul>	<ul style="list-style-type: none"> <li>• Can you see YouTube videos?</li> <li>• Can you see your Echo Show camera on the cloud app?</li> <li>• Can it connect to music speakers?</li> </ul>
Sovaro Luxury Cooler		<ul style="list-style-type: none"> <li>• Will this thing make me fly? It seems due to the price that it has to do something special</li> <li>• Which organ should I sell to finance this ice box?</li> <li>• Just how insecure do you have to be to buy one?</li> </ul>	<ul style="list-style-type: none"> <li>• Will this fit in the trunk of my Lambo?</li> <li>• Where do you plug it in?</li> <li>• What is the country of origin?</li> </ul>
Hutzler 571 Banana Slicer		<ul style="list-style-type: none"> <li>• I set it down in my kitchen, my bananas have stopped talking to me. What now?</li> <li>• What if the banana bends the other direction?</li> <li>• Is there a model for left-handed people?</li> </ul>	<ul style="list-style-type: none"> <li>• Can this be used on cucumbers?</li> <li>• Does it only come in yellow?</li> <li>• Seriously, though - does this thing really work?</li> </ul>

Previous studies have used unique textual properties for assisting humor detection [39]. One main property is *incongruity* [2], where humor arises from a surprising or inconsistent situation, opposition, and other forms of apparent contradiction. While in previous work incongruity was detected within context-free sentences [38], detecting it in-context has not yet been tackled. In our framework, we measure incongruity according to the propensity of the question from the product (e.g., “Could Echo cook breakfast?”).

Another useful property for humor detection is *subjectivity* [16, 37, 39], as reflected by the emotion expressed in the text. For example, the question “Will this Swiss knife make me happy?” expresses subjective suspicion in the product capabilities, as detailed by the manufacturer. Utilizing subjectivity for humor detection has not attracted much attention in deep learning setting. In this work we study how text subjectivity can be utilized in our deep-learning framework.

A major technical challenge in humor detection is *domain bias* which may happen when a classifier is trained to detect humor in text according to its domain. This may happen when positive and negative examples are sampled from multiple domains with different distributions over the data. Previous studies have either ignored the domain bias, or eliminated it by selecting training examples from the same domain [20, 39]. In the PQA setting, domain bias appears when some products attract more humorous questions than others, consequently, a classifier may be trained to identify such products, rather than identifying humorous questions. The canned unicorn meat box, presented in Figure 1, is a good example

for a product that attracts many humorous questions. Other common types of humor-attracting products are adult toys, peculiar and bizarre products, and ridiculously expensive products.

In this work we develop a deep-learning framework for detecting humorous questions in PQA systems, incorporating information gathered from the question and the associated product. Our learning framework integrates two additional informative sources: one captures the incongruity between the question and the product, and the second captures the question’s subjectivity level. We compare the performance of our learning system with several baselines, using a large dataset of product-related humorous questions, collected from a commercial PQA system. In order to eliminate the product bias we experiment with two balanced datasets, both sharing the same set of humorous questions, which are different in the way negative examples are selected. For the biased dataset, negative examples are selected at random from the entire population of questions. For the unbiased dataset, a negative example is selected at random, for each positive question, from the set of non-humorous questions that match the question’s matching product. As expected, a classifier trained over the unbiased dataset, achieves lower accuracy than a classifier trained over the biased dataset. However, the unbiased classifier excels in the task of differentiating between humorous and non-humorous questions that match the same product.

The main contributions of our work are as follows:

- We collected a first-of-a-kind dataset of humorous and non-humorous questions in the PQA domain.

- We developed a framework for detecting humorous questions in PQA systems, while considering two main humor related properties: *Incongruity* and *Subjectivity*.
- We demonstrated the existence of product bias in PQA domains, and the necessity of bias elimination for the humor detection task.
- We experiment with our classifiers over the biased and unbiased datasets, demonstrating classification accuracy of 90.8% of the biased classifier, and 84.4% of the unbiased one, a relative improvement of 18.3% and 5.4% over baseline methods.

The rest of the paper is organized as follows: Section 2 surveys existing work. Section 3 formalizes our research problem, and describes our learning framework and two additional humor related attributes – incongruity and subjectivity. In Section 4 we present the two datasets used for our experiments, with and without product bias. Section 5 experiments with our humor detection system and compares it with several baselines, and presents potential usage of this work. Finally, Section 6 concludes our work.

## 2 Related Work

We next review related work in the context of PQA-related tasks, and deep learning methods for humor detection.

### 2.1 PQA-Related Tasks

PQA introduces new challenges due to the manifestation of questions in the context of a specific product, including product details and product reviews. McAuley and Yang [21], followed by others [28, 35], have predicted the answer for a product question by utilizing product reviews. Other research challenges relate to evaluating answers quality [33] and answer ranking [1, 10, 27], extensively studied in the CQA domain but not yet in PQA. Detecting humorous questions can contribute to such tasks, for example, by considering humor as an additional question attribute that may affect these tasks, as well as ranking the product questions according to their humorous level. Another important service for PQA users is to mark humorous questions, thus alerting users about their flippancy; such questions are not needed to be answered too seriously, or that their answers should be taken with a grain of salt.

### 2.2 Humor Detection

Yang et al. [39] detected humor in puns and one-liners. Their seminal work introduces the foundations of humor detection that provide the baselines for our study. First, they presented the semantic structures in humorous content including: (a) Incongruity, (b) Ambiguity, (c) Interpersonal effect (including sentiment and subjectivity), and (d) Phonetic style. We further discuss incongruity and subjectivity in the following. For each semantic structure they designed ad-hoc representative features and trained (non-deep) classifiers for humor recognition. They experimented with the four structures mentioned above and found that incongruity outperforms other structures in contributing for humor recognition.

*Incongruity* has been acknowledged as one of the main technique for generating humorous text [34]. According to the incongruity theory, as published by Berlyne [2] (who cites Beattie, 1776), “*Laughter arises from the view of two or more inconsistent, unsuitable, or incongruous parts or circumstances, considered as united in one complex*

*object or assemblage*”. Previous studies offered several approaches to measure incongruity between two pieces of text. For example, Yang et al. [39] calculated the maximum and the minimum distance between word embedding pairs in two pieces of text as a signal for incongruity.

*Subjectivity* was introduced as another property of humorous text in several studies [16, 37]. In many cases, a subjective opinion, e.g., criticism, is expressed by humor. Karimi and Azadeh [17] detected text subjectivity using two language-models, one trained over objective data while the other over subjective data. Specifically, they used movie reviews as a subjective text, and movie summaries as an objective text. The distances between the text’s language model to the objective and subjective language models were used to determine the subjectivity level. Inspired by this direction, we detect subjectivity in questions by learning a subjective language model from product reviews, and an objective one from product descriptions.

### 2.3 Domain Bias

Another research challenge introduced by Yang et al. [39] is the existence of domain bias in humor detection. In their study, positive examples were taken from dedicated datasets including puns-of-the-day and one-liners, while negative samples were randomly sampled from other sources, including AP News, New York Times, Yahoo! Answer and Proverb. This raises the question whether their proposed method detects the humor itself or the text’s domain. In order to reduce the domain bias, Yang et al. selected, as negative examples, sentences that are similar to the positive examples. They did it by selecting only sentences whose words are found in the vocabulary of the positive examples, and their length is similar to the length of the positive examples. They did not report on the effect of bias elimination on humor classification.

For some domains, eliminating the domain bias is natural. For example, Lee et al. [20] predicted audience laugh during a TED talk. The negative and the positive examples were taken from the same talk hence there is no (talk) bias. The accuracy of their classifier was only 53%, indicating the difficulty in detecting humor from unbiased training data. Similarly, Bertero and Fung [3], extracted positive and negative humorous examples from one television show script, hence their data does not suffer from a (series) bias. They trained an LSTM-based framework, utilizing previous sentences as context, achieving F1 score of 62.9% and accuracy of 70% in identifying humorous content over a non balanced dataset, however, their model was not tested on other shows.

Recently, additional studies have applied deep learning methods for humor detection. Chen and Lee [5] developed a deep-learning framework for monitoring humor in Ted talks and in puns. They compared orthodox classification methods to a CNN method, showing that CNN outperforms the baselines. Their learning framework, which does not utilize concrete humor features, has reached an accuracy of 58.9% over the Ted dataset [20]. Chen and Soo [6] showed that a CNN classifier, based on the text only, outperforms a classifier that is based on the manually created features proposed by Yang et al. [39]. They experimented with various datasets including pun-of-the-day, one-liners [22] and a Chinese dataset created especially for this task. Weller and Seppi [36] applied a transformer architecture

to detect humor within Reddit Jokes threads and punch-of-the-day dataset, showing significant improvement over other deep learning methods for humor detection.

## 2.4 Other Humor Detection-Related Tasks

Several humor-related studies mentioned that humor is not necessarily funny or even aspire to be funny [7, 13]. While humor expresses an objective intention of the question’s author, funny is a subjective interpretation of the reader. Gulas and Weinberger [13] have studied humor in advertising. They pointed out ethnic boundaries that make jokes untranslatable sometimes. Chairó [7] analyzed linguistic and cultural barriers in humor translation, revealing cultural differences that challenge readers to understand humor, or to find it funny.

*Sarcasm*, a specific type of humor, was detected in several studies. Peled and Reichart [29] created a training dataset using tagged tweets, and applied monolingual machine translation algorithm to alter sarcastic tweets into non-sarcastic ones. Xiong et al. [38] detected sarcasm over context-free sarcastic sentences extracted from social networks such as Reddit and Twitter. They proposed a novel model for incongruity detection based on identifying contradiction between word meanings, using an attention model of word co-occurrences.

De Oliveira and Lainez Rodrigo [8] detected humor over a dataset of Yelp reviews using deep learning methods. They pointed out the need of a humorous dataset with varying humor attributes and varying sources. Potash et al. presented a SemEval dataset [31] for humor detection based on responding tweets, sent as part of the TV show *@midnight*, in which participants (usually comedians or actors) are requested to post funny tweets with respect to a given topic. The label (how funny a tweet is) was determined based on the tweet’s comments. The dataset was published together with two related tasks: 1) given two tweets, decide who is funnier, and 2) rank a list of tweets according to their humor level. These tasks are similar to the humor-based question ranking task in PQA domains. Miller et al. [25] presented another SemEval dataset for detecting English puns. They focused on context-free humor detection and introduced three subtasks: pun detection, pun location, and pun interpretation.

Hasan et al. [14] presented a multimodal dataset of TED talks consisting of textual, visual, and acoustic features, to detect humorous segments within the talks. They trained a deep learning network to process all features together. They concluded that the context and the punchline are the most important components in text for humor detection. This is correlated with the general understanding about the importance of incongruity for humor detection, as was shown in previous studies, and is also supported by our work of detecting humorous questions in PQA domains.

## 3 Research Problem and Detection Framework

In this Section we state our research problem—detecting humorous questions in PQA domains. Then, we describe our deep-learning framework for handling this problem.

**Problem statement:** Given a product-related question, detect if it is humorous or not, in the context of its matching product.

## 3.1 Humor Detection Framework

We describe our humor detection framework and its internal components. We begin with a short review of deep learning methods as they are an integral part of our framework. Then, we describe in details the framework’s different components.

**3.1.1 Deep Learning Methods.** In recent years artificial neural networks (ANN) have emerged as popular frameworks for text classification. A *Convolution Neural Network (CNN)* is an ANN that includes a convolution layer with nonlinear activation functions [18]. The input to a CNN network is an embedding layer, typically using a pre-trained embedding framework such as Glove [30], or FastText [23]. The output of the CNN can be converted by a softmax function into a distribution over the predicted classes. A *Recurrent Neural Network (RNN)* is another type of ANN [24]. Unlike feed-forward neural networks RNNs store an internal memory to process a sequence of inputs. An LSTM network is an RNN that is based on LSTM units [15]. Its main advantage over classic RNN is that an LSTM network tackles the vanishing or exploding gradient problem in a sequence of text due to the LSTM unit properties.

**3.1.2 Framework Components.** A sketch of our framework is presented in Figure 2. The input for the framework is a question and a product title, and the output is the humor detection flag. Our framework includes two pre-trained modules that capture *incongruity* and *subjectivity* as these were recognized, by previous work, as major indicators for humor in text. In addition to these modules, the framework includes two representation models, denoted by *Q-Rep* and *T-Rep*, that capture features from the question and the associated product title, and a final classifier that processes the output of all other components.

The product title is fed into the *T-Rep* model and into the incongruity module. The question is fed into *Q-Rep* model, into the incongruity module, and into the subjectivity module. The outputs of the different modules are then concatenated together into a single vector, including the deep embedding vectors from representation models (*T-Rep* and *Q-Rep*), the output vector from the subjectivity module, and the incongruity module score. This final vector is then provided as an input to a fully connected layer which outputs the humorous decision.

The incongruity and subjectivity modules are trained on auxiliary datasets, and their internal weights are kept fixed during the humor detection training phase. In contrast, the *T-Rep* and *Q-Rep* models, and the final layer, are trained during the training phase of the humor detection network.

***T-Rep* and *Q-Rep*.** In order to represent the product title and the question in our model, we attach an LSTM (or CNN) network that captures their inner features. The deep embedding vector, i.e., the last hidden layer of LSTM, or the flatten layer of CNN, is used as their vector representation.

**3.1.3 Incongruity.** As detailed in the Section 2, incongruity describes a situation where humor arises from inconsistent, unsuitable or incongruous parts of the situation [2, 34]. In our case, incongruity refers to discrepancy of a question from its matching product. A question that is more probable to appear together with the product (with low incongruity) is less likely to be humorous, while

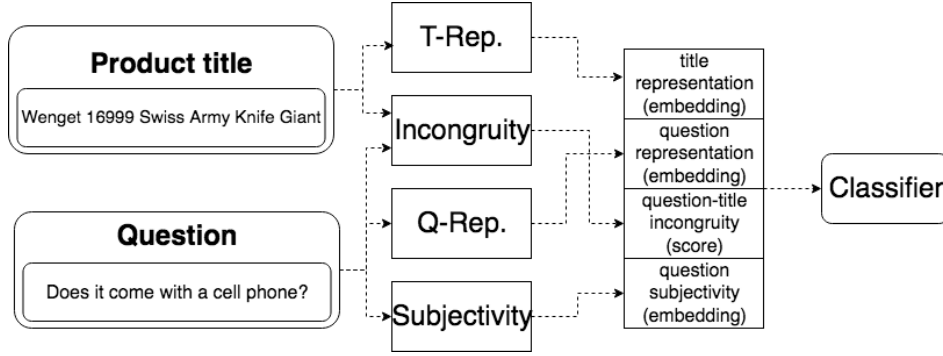


Figure 2: Humor Detection Framework

in contrast, implausible association (high incongruity) indicates humorous intention. For example, the association between cleaning the dishes with a dish washer is probable, while association between cleaning dishes and a computer is unlikely.

A common approach to capture probable appearance of two pieces of texts is by using a *Siamese* network [26]. Technically, A Siamese model shares the weights of two LSTM networks, and by doing so, it enforces the same representative features for both input texts. The training process applies a loss function bringing closer similar texts and distancing dissimilar ones. In our case, the network gets as input a (question, product title) pair, and feeds them simultaneously into two copies of the same LSTM network, enforcing similar representations for positive example pairs, and dissimilar representations for negative example pairs.

Once the network is optimized, its weights are fixed and the model is used ‘as-is’ by the overall framework. Figure 3 presents our LSTM-based incongruity network<sup>3</sup>:  $q_1, \dots, q_n$  and  $t_1, \dots, t_m$  represents word embedding of the question and the title respectively;  $h_t$  and  $h_q$  denote the last hidden states of the two LSTM networks. We use the same loss function as described in [26], measuring  $L_1$  distance between the hidden vectors of the question and the title:

$$\text{loss}(q, t) = \exp(-||h_q, h_t||_1).$$

To train the incongruity network, we extract 100K random (product title, question) pairs from the general PQA corpus as positive examples, since most questions are plausible for their matching product. 100K Negative examples are collected by mixing (randomly selected) questions and products, creating artificial improbable (product, question) pairs that are unlikely to appear together. Table 2 reveals some examples for questions identified by the trained network as having high incongruity with respect to their matching product.

**3.1.4 Subjectivity.** Humorous textual content usually expresses an interpersonal effect [39] like sentiment or subjective opinion. Consider the question from Table 1, “*Will this thing make me fly? It seems due to the price that it has to do something special*”. This question was asked in the context of a luxury cooler. The asker clearly expresses a negative sentiment about the product price.

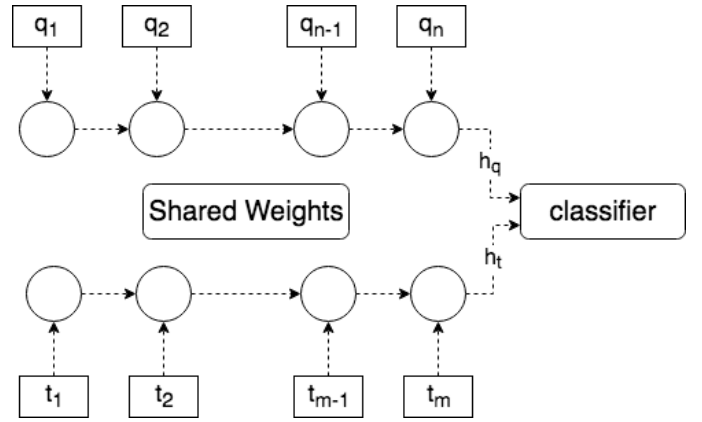


Figure 3: Incongruity network

Table 2: Examples for questions identified as having high incongruity with respect to their matching product.

Product Title	Question
Nokia 3310 Cell Phone	Does it stop bullets?
Apple iPhone X	Does it wash the dishes and take out the trash?
Pebble Time Steel Smart-watch	Will it cook breakfast and make coffee?
Full Face Fleece Warm Winter Sports Mask	Will it give the wearer any mystical ninja-like abilities?

*Sentiment Polarity Analysis* was vastly learned in recent years [4, 12, 40] with the goal of correctly classifying positive and negative sentiment in textual content. In order to learn the question’s sentiment polarity, we trained an LSTM model based on the Blitzer’s dataset [4] consisting of Amazon product reviews. The sentiment of each review is labeled according to its associated human star rating. However, the sentiment polarity signal was found as unhelpful in detecting humorous questions. Our initial examination showed an accuracy of 52% in humor detection based on this signal. A potential explanation is that humorous questions can either by positive (“will

<sup>3</sup>We also experimented with a CNN based Siamese network which performed worse.

this thing make me fly?") or negative ("this poor item will make you cry!"). We therefore abandon the sentiment polarity signal and focused on subjectivity.

Following previous studies on the role of subjectivity for humor detection [16, 39], we assume that humorous questions tend to be subjective, while non-humorous questions are more likely to represent an objective point-of-view. Inspired by Karimi et al. [17], we used product reviews as examples for subjective text and product descriptions as examples for objective text. We trained an LSTM network based on a randomly curated set of 20K product reviews and a same number of product descriptions. During inference time, the subjectivity network takes as an input the question’s embedded terms and outputs the final LSTM hidden vector as a subjectivity feature for the overall framework. This module can be used to detect humor directly, independently of the rest of the framework, by training a classifier that gets as input the subjectivity vector and outputs a humor flag (for more details see Section 5.4). Table 3 shows some questions identified by this network as highly subjective.

**Table 3: Examples for questions identified as highly subjective.**

<b>Subjective Questions:</b>
- Is this glue strong enough to hold my parents marriage together?
- Will this [product] make me happy and take away all my problems back?
- If Harambe [the Gorila] had slept on this [product] would he still be alive? Asking because I’m an angry man.

## 4 Dataset

Our evaluation framework is based on two novel datasets of product related humorous questions, associated with their matching product titles. The two datasets, created especially for this work, share the same positive humorous questions and differ with respect to the negative non-humorous examples<sup>4</sup>. We next describe the humorous questions annotation process, and provide some general dataset statistics.

### 4.1 Data Collection

Our data is based on a large set of humorous questions extracted from a commercial PQA system. We re-validated the humor of these questions using human annotators. An example of the annotation task page is presented in Figure 4. Annotators were presented with a question and the associated product title and image, and were asked to classify whether the question is humorous. Annotations were done using the Figure-Eight framework<sup>5</sup>; each question was judged by at least three and up to seven annotators, until reaching an agreement level of at least 70%. Agreement level per question is measured by the portion of annotators who agreed on the label. The high agreement level among annotators (Fleiss’ kappa [11] of 0.67 among the first three annotators who judged the question independently, and an average agreement level of 89.5% among

Question : can you travel back in time wearing this watch?

Description : Apple Watch Series 2 42mm  
(Gold Aluminum Case Midnight Blue Sport Band)  
MQ152LL/A

The question is: (required)

- ☐ Not entirely in English  
☐ Not humorous  
☒ Humorous



**Figure 4: Annotation task page as presented to annotators**

all annotators) reveals that people tend to highly agree whether a question is humorous or not.

### 4.2 Unbiasing the data

Previous studies have shown that it is important to select both positive and negative examples from the same domain, otherwise, classifiers are prone to detect the domain rather than the humor itself [3, 20, 39]. In PQA systems, domain bias exists since some products attract more humorous questions than others. We denote this bias as *product bias*. For example, we found the bizarre canned unicorn meat box, presented in Figure 1, to be associated with many humorous questions, probably due to its peculiar nature. The specific setting of PQA enables us to eliminate the product bias by selecting the training instances, both humorous and non humorous questions, from the same product. For example, the Nintendo Switch presented in Table 1 is associated with both humorous (e.g., "will the switch cure cancer?") and non-humorous questions ("how much money will the system cost?").

To examine the product bias effect on humor detection we created two *balanced* datasets (equal amount of humorous and non-humorous questions). Both datasets contain the full set of humorous questions as positive examples. They differ in the way negative examples were collected.

The first, *biased* dataset, consists of the same number of humorous and non-humorous questions. The non-humorous questions were selected at random from the entire population of product questions.

For the second, *unbiased* dataset, for each humorous question, we randomly selected a non-humorous question that matches the same product. This reduces product bias since the same subset of humor attracting products is covered by both positive and negative example sets, hence it has a similar effect on both. This is in contrast to the *biased* dataset, where the number of humor-attracting products which are covered by the positive set, is larger than those covered by the negative set. In both datasets, we validated the non-humorous of the negative examples by manual annotations.

### 4.3 Data Analysis

Table 4 presents the number of questions and the covered products in each dataset. The datasets consist of the same number of questions, 19,142, since they share the same set of humorous questions and each holds the same number of non-humorous questions. Both datasets were split to train, validation, and test sets, consisting of

<sup>4</sup>Our dataset is available via: <https://registry.opendata.aws/humor-detection/>

<sup>5</sup>[www.figure-eight.com](http://www.figure-eight.com)



60%, 20% and 20% of the questions respectively. There is no intersection of products between the splits, that is, no product appears in more than one split.

The overall number of covered products by the biased dataset, 12,701, is more than twice than those covered by the unbiased dataset, 6,020, as for each humorous question in the biased dataset another question is selected randomly from the whole question population. When counting the number of humorous questions per product, 84.8% of the products are associated with a single humorous question, 6.44%, 3.19% and 2.53% are associated with two, three and four questions. Only 3.04% of the products are associated with five or more questions. Note that this does not apply to the distribution of humorous questions in the overall corpus but only displays the distribution in our datasets.

**Table 4: unbiased and biased datasets statistics.**

Dataset	Category	All	Train	Validation	Test
<i>biased</i>	# of questions	19142	11484	3828	3830
	# of products	12701	6821	2912	2968
<i>unbiased</i>	# of questions	19142	11484	3828	3830
	# of products	6020	2671	1583	1766

Figure 5 shows the (normalized) distributions of humorous and non-humorous questions across product categories in our data. It is clear that the distribution of humorous questions differs significantly from the non-humorous questions distribution. In particular, frequent categories in our data (e.g. 'pc', 'electronics', 'wireless') do not contribute many humorous questions. As can be expected, many humorous questions come from the 'toy' category which include adult toys.

## 5 Evaluation

The humor detection evaluation framework was utilized to answer the following research questions:

- What is the performance of the overall humor detection framework?
- What is the impact of product bias on system performance?
- What is the performance of the sub-tasks (incongruity and subjectivity) and their contribution to the overall framework performance?

In order to answer these questions, we conduct the following experiments. First, we quantify the effect of the product bias. Then we evaluate the overall performance of the framework, by comparing it to baseline methods. We conduct an ablation test, using different configurations of our classifier, to estimate the contribution of incongruity and subjectivity. Finally, we evaluate independently the performance of the incongruity and subjectivity modules.

### 5.1 Methods

*Baseline Methods.* We consider the following baseline methods for humor detection:

- *Linear classifiers.* Naive Bayes and Logistic Regression [32] using unigrams and bigrams which appear at least five times in the data.

- *Basic Deep learning Classifiers.* LSTM [15] and CNN [19] considering only question's text, denoted as  $LSTM_Q$  and  $CNN_Q$ . CNN effectiveness was proven for humor detection by [6] and it outperforms decision trees as shown by [39].

*Proposed Methods.* We consider the following novel methods:

- *Question and Product Title.* Denoted by  $LSTM_{Q+T}$  and  $CNN_{Q+T}$ . These are LSTM and CNN networks that take as input the concatenation of  $Q$ -Rep and  $T$ -Rep, the representations of the question text and the product title.
- *With incongruity.* Consider  $T$ -Rep and  $Q$ -Rep, jointly with the incongruity score. Denoted by  $LSTM_{INC_{Q+T}}$  and  $CNN_{INC_{Q+T}}$ .
- *With subjectivity.* Consider  $T$ -Rep and  $Q$ -Rep, jointly with the subjectivity embedding vector. Denoted by  $LSTM_{SUB_{Q+T}}$  and  $CNN_{SUB_{Q+T}}$ .
- *Complete Framework.* Consider  $T$ -Rep and  $Q$ -Rep, jointly with the incongruity score and the subjectivity embedding vector. Denoted by  $LSTM_{INC\_SUB_{Q+T}}$  and  $CNN_{INC\_SUB_{Q+T}}$ .

In all the modules of our framework ( $T$ -Rep,  $Q$ -Rep, *incongruity* and *subjectivity*) the raw text is converted into embedding using a Wikipedia based, Fasttext pre-trained embedding [23]. We used an embedding dimension of 300 and preserved only the top 10K most frequent words. We trained CNN using kernel size of 3 (tri-grams). Hyper parameter tuning adjusted the filters (a sliding window over the network input) considering windows of size 32, 64 and 128. LSTM was optimized considering 32, 64 and 128 hidden states. All the models were trained over EC2 G3 8x-large instances<sup>6</sup> consisting of 2 GPU units, each assigned with 8 GB of memory. Code was written in Python using the Keras package<sup>7</sup>.

### 5.2 Product Bias Effect

To demonstrate the product bias, we examine the ability to identify a humorous question based only on the product title. In order to do so, we trained a deep learning network (CNN and LSTM) replacing each question with its matching product's title. If no bias exists, title should not imply any signal and classification should demonstrate (close to) random performance. However, if the title does enclose a signal, the network should learn how to use it for humor detection.

In the unbiased dataset, after replacing questions with titles, each title appears the same number of times – half of the times marked as humorous and half of the times as non-humorous. As expected, prediction accuracy of the trained models was exactly 50%.

The bias was revealed while using the biased dataset, where questions were replaced with titles as done before. Since there is negligible overlap at the product level, each title is associated either with a humorous or with a non-humorous label. The accuracy of the LSTM model over this dataset was 73.99 and the accuracy of the CNN model was 72.25. It can be clearly seen that the models manage to learn (to some extent) the humorous label using only product title, which implies that titles do convey a signal. Note that we could measure the bias since we had the product context in hand, which is unavailable when humor is detected over context-free text (e.g., a single sentence). To the best of our knowledge, this is the

<sup>6</sup><https://aws.amazon.com/ec2/instance-types/g3/>

<sup>7</sup><https://keras.io/>

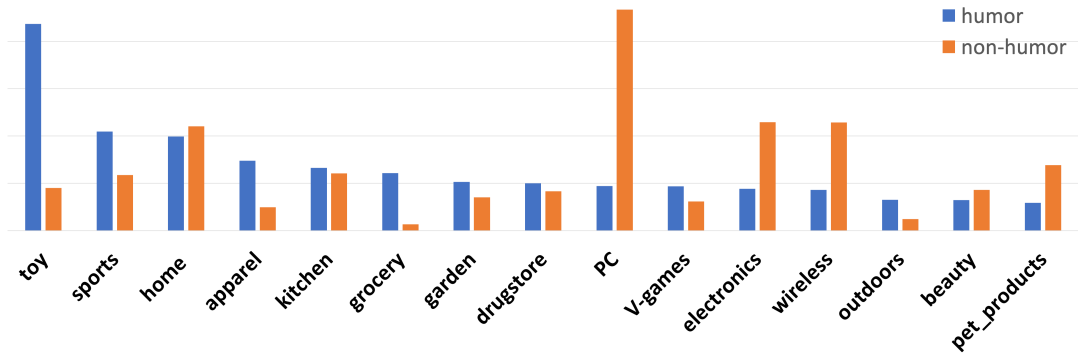


Figure 5: Humorous and non-humorous questions distributions across categories.

**Table 5: Accuracy of humor detectors over the two datasets. Statistically significance improvement with respect to baseline methods is marked by “\*”, using McNemar test with p-value < 0.05 [9].**

Configuration	Method	unbiased	biased
Baseline Methods	Logistic Regression	82.11	87.21
	Naive Bayes	81.64	87.62
	$LSTM_Q$	83.52	88.59
	$CNN_Q$	83.26	88.69
Partial Configuration	$LSTM_{Q+T}$	83.42	89.74
	$CNN_{Q+T}$	83.63	89.58
	$LSTM_{INC_{Q+T}}$	83.58	89.8
	$CNN_{INC_{Q+T}}$	83.26	90.34
	$LSTM_{SUB_{Q+T}}$	83.81	90.08
	$CNN_{SUB_{Q+T}}$	83.71	90.23
Complete Framework	$LSTM_{INC\_SUB_{Q+T}}$	<b>84.41*</b>	90.26
	$CNN_{INC\_SUB_{Q+T}}$	84.13	<b>90.76*</b>

first time product bias is explicitly quantified in the literature, in the context of humor detection.

### 5.3 Methods Performance

Table 5 presents classification accuracy of all methods over the two datasets. As expected, the results reveal that the unbiased dataset provides a harder-to-learn setting. The performance of the baseline methods confirms that deep networks outperform orthodox classification methods [6].

The complete framework reduces the error with respect to the best baseline method in both datasets. In the biased dataset relative error is reduced by  $-18.3\%$  (90.76% vs 88.69%) and in the unbiased dataset by  $-5.4\%$  (84.41% vs 83.52%). The difference was found to be statistically significant by McNemar’s test, compared to the strongest baseline in each dataset (i.e.,  $LSTM_Q$  for the unbiased dataset and  $CNN_Q$  for the biased dataset).

In order to study the contribution of the different framework components, we used partial configurations as presented in Table 5. For the biased dataset, using  $Q$ -Rep and  $T$ -Rep only (without incongruity and subjectivity) has a positive effect on performance,

compared to the baselines. Adding *incongruity* and *subjectivity* independently, further improved the performance. In comparison, the effect of each module is marginal over the unbiased dataset, while their overall integration contributes 5.4% in relative error reduction.

### 5.4 Subtasks Evaluation.

We evaluate the independent performance of the incongruity and subjectivity modules over the two datasets. Since both modules were trained as auxiliary modules, we evaluated their humor detection capability independently from the entire framework.

The input of the subjectivity module is a question text, while the input of the incongruity module is a pair of a question and its matching product title. The output is the subjectivity vector, and the incongruity score, respectively. To evaluate the performance of the subjectivity module, we appended a dense layer on top of the module, which receives the subjectivity vector as input, and takes a humor decision. For incongruity, we learned a threshold-based decision rule on the module’s incongruity score, marking a question as humorous when the score is below 0.5. Evaluation was done by comparing each module’s prediction with the humorous annotated label.

The accuracy results are depicted in Table 6. The *subjectivity* module presents a relatively good performance over the two datasets, supporting the assumption that it is a good indicative for humor detection. Incongruity also performs reasonably well, though a bit weaker. A difference in performance can be detected in favor of the biased dataset, probably due to the product bias.

**Table 6: Accuracy of the *incongruity* and *subjectivity* modules for the humor detection task.**

Subtask	unbiased	biased
Subjectivity	63.2	66
Incongruity	54.25	60.4

**5.4.1 Deeper analysis.** In order to better understand these results we deep dive into both modules. We calculated the average subjectivity score (the output of the dense layer on top of the subjectivity



network), for several sets: humorous questions, with average subjectivity score of 0.36, non-humorous from the unbiased dataset, with a score of 0.29 and non-humorous from the biased dataset, with score of 0.31. The higher average subjectivity score of humorous questions attests our assumption that humorous questions are more subjective than non-humorous questions.

We further investigate the humor detection performance of the *incongruity* module. When incongruity score is fed to the overall humor detection system as a feature (Table 5), or independently (Table 6), it is less effective in the unbiased setting compared to the biased one. To further explore the reasons for this gap, we fed (question, product title) pairs from both datasets into the incongruity network, to predict how probable the question is with respect to its product. Since all pairs in our datasets originally appear together in the PQA system, their incongruity label is negative<sup>8</sup>. Therefore, we calculate recall for each label, i.e., measuring the portion of humorous questions that were correctly classified, and similarly for non-humorous questions. Results are presented in Table 7.

**Table 7: Recall of *incongruity* module over the two datasets.**

Label	unbiased	biased
Humorous	52.3	52.3
Non-Humorous	65	71.8

In both datasets, we see that the humorous examples are less probable (hence with higher incongruity) compared to the non-humorous examples, in agreement with our hypothesis. We also see that non-humorous examples in the unbiased dataset are less probable than non-humorous questions in the the biased dataset. This may be explained because any non-humorous question in the unbiased dataset, matches a product attached with at least one humorous question, hence, it is likely that many of these products are humor-attracting. For such a humor-attracting product, some matching questions may be incongruous, including non-humorous ones. For example, the non-humorous question “Is this stuff safe to spray directly on your skin?” has high incongruity with its peculiar product ‘Pheromones for man — attract women’.

Moreover, we can further see that the gap in performance between non-humorous and humorous examples in the biased dataset (19.5%) is higher than in the unbiased dataset (12.7%); this difference explains the better performance of incongruity over the biased dataset because of the better differentiation between humorous and non-humorous examples in this setting.

## 5.5 Usage

We consider two use-cases where humor detection can be applied in PQA systems. In the first use-case, we are given a question selected at random from the whole corpus. This can represent, for example, humor detection of a new question upon its publication. In the second use-case, we are given a product and need to classify (or rank) its questions with respect to their humor. This can represent, for example, a use-case of identifying humor-attracting products.

<sup>8</sup>We would like to emphasize that in the dataset used to train the incongruity network, questions and products were randomly mixed as training examples with a positive incongruity label (See Subsection 3.1.3).

These two use-cases can be served by two different classifiers, one is optimized over the biased dataset and the other is optimized over the unbiased dataset, denoted by *biasedQClassifier* and *unbiasedQClassifier* respectively. The *biasedQClassifier* is optimized to detect humorous questions while utilizing the product bias while *unbiasedQClassifier* is optimized to detect humorous questions when product bias is eliminated.

In order to verify that indeed each classifier was optimized for its own use-case, we measured the accuracy of each classifier over the test set of the other dataset. Results are presented in Table 8. It can be seen that the *biasedQClassifier* outperforms *unbiasedQClassifier* on the biased dataset, showing that it indeed utilizes the bias to improve detection. Similarly, *unbiasedQClassifier* outperforms *biasedQClassifier* over the unbiased dataset, showing that *unbiasedQClassifier* detects humorous in absence of product bias.

**Table 8: Classifiers accuracy over the two datasets**

Dataset	<i>unbiasedQClassifier</i>	<i>biasedQClassifier</i>
biased	82.85	90.76
unbiased	84.41	76.21

## 6 Conclusion

In this study we presented a deep-learning framework for detecting humorous questions in PQA systems. This detection is beneficial due the popularity of such questions among customers, hence, by boosting them on the products’ detail pages we can expect a better user engagement. Our framework considers information from the question and the product title and integrates it with two pre-trained auxiliary modules: incongruity and subjectivity.

The second challenge studied in this work deals with product bias in PQA data. Experiments reveal that we reach an accuracy of 90.76% and 84.41% over the datasets, with and without the bias respectively. This is an improvement of 18.3% and 5.4% in relative error reduction over baseline methods described in previous studies. We also experimented with two classifiers, each trained independently over one of the datasets. Each of the classifiers can be used in a different use case, therefore, both are required in different scenarios.

Our study shows that product bias is derived from the product’s own characteristics, and is typically related to peculiar and bizarre products, such as adult toys, highly expensive products, and products with unexpected functionality. Detecting products that attract humorous questions is an interesting research challenge which we leave for future work.

While the humor detection accuracy of our classifier exceeds 90% over the biased data, there is still a lot of room for improvement. One direction is to investigate humor detection in bias-free data, where detection accuracy is relatively lower. Another direction is studying humor detection in real-world setting, over non-balanced data, where the prevalence of humorous questions is much smaller than non-humorous ones.

Another potential direction is applying external knowledge for better understanding the question’s context. Consider, for example,

the question from Table 1, asked about the Nintendo Switch; “*What if the princess wants to be with Bowser and Mario keeps kidnapping her?*”. The context is that Mario keeps saving the princess in this game, maybe against her will, as the question hints. Understanding the context is essential for detecting the humor in such questions. Furthermore, detecting humorous answers, is a very interesting research direction, from research and practical perspectives, as well as for humor lovers who can enjoy good humor while looking for an answer to their question.

## References

- [1] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM*. ACM.
- [2] Daniel E Berlyne. 1969. Laughter, humor, and play. *The handbook of social psychology* 3, 2 (1969).
- [3] Dario Bertero and Pascale Fung. 2016. A Long Short-Term Memory Framework for Predicting Humor in Dialogues. In *NAACL HLT The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA*.
- [4] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic*.
- [5] Lei Chen and Chong Min Lee. 2017. Convolutional Neural Network for Humor Recognition. CoRR abs/1702.02584 (2017). arXiv:1702.02584
- [6] Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 113–117.
- [7] Delia Chiaro. 2011. Comic takeover or comic makeover?: Notes on humour-translating, translation and (un) translatability. In *The pragmatics of humour across discourse domains*. John Benjamins, 365–378.
- [8] Luke De Oliveira and Alfredo L Rodrigo. [n.d.]. *Humor detection in Yelp reviews*. Technical Report. Stanford Institute for Computational and Mathematical Engineering, California, USA.
- [9] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 1383–1392.
- [10] Minwei Feng, Bing Xiang, Michael R. Glass, Lidian Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*.
- [11] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.
- [12] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA*. Omnipress, 513–520.
- [13] Charles S Gulas and Marc G Weinberger. 2010. That’s not funny here: Humorous advertising across boundaries. *Translation, Humour and the Media: Translation and Humour* 2 (2010), 17–34.
- [14] Md. Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md. Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*.
- [15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [16] Manan Jain. 2017. *Humor Detection*. Ph.D. Dissertation. Harrisburg University of Science and Technology.
- [17] Samaneh Karimi and Azadeh Shakery. 2017. A language-model-based approach for subjectivity detection. *Journal of Information Science* 43, 3 (2017), 356–377.
- [18] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, ACL*, 1746–1751.
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, Nevada, United States*. 1106–1114.
- [20] Chong Min Lee, Su-Youn Yoon, and Lei Chen. 2016. Can We Make Computers Laugh at Talks?. In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media, PEOPLES@COLING 2016, Osaka, Japan*. The COLING 2016 Organizing Committee, 173–181.
- [21] Julian J. McAuley and Alex Yang. 2016. Addressing Complex and Subjective Product-Related Queries with Customer Reviews. In *Proceedings of the 25th International Conference on World Wide Web, WWW*. ACM, 625–635.
- [22] Rada Mihalcea and Carlo Strapparava. 2005. Making Computers Laugh: Investigations in Automatic Humor Recognition. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada*.
- [23] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhresch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan*.
- [24] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan*. ISCA, 1045–1048.
- [25] Tristan Miller, Christian Hemp, and Iryna Gurevych. 2017. SemEval-2017 Task 7: Detection and Interpretation of English Puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 58–68.
- [26] Jonas Mueller and Aditya Thyagarajan. 2016. Siamese Recurrent Architectures for Learning Sentence Similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, Phoenix, Arizona, USA*. AAAI Press, 2786–2792.
- [27] Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 Task 3: Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada*. Association for Computational Linguistics, 27–48.
- [28] Jianmo Ni, Zachary C. Lipton, Sharad Vikram, and Julian J. McAuley. 2017. Estimating Reactions and Recommending Products with Generative Models of Reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP, Asian Federation of Natural Language Processing*, 783–791.
- [29] Lotem Peled and Roi Reichart. 2017. Sarcasm SIGN: Interpreting Sarcasm with Sentiment Based Monolingual Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada*. Association for Computational Linguistics, 1690–1700.
- [30] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, ACL*, 1532–1543.
- [31] 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)*. Association for Computational Linguistics, Vancouver, Canada, 49–57.
- [32] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.
- [33] Chirag Shah and Jeffrey Pomerantz. 2010. Evaluating and predicting answer quality in community QA. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*. ACM, 411–418.
- [34] Oliviero Stock and Carlo Strapparava. 2003. Getting Serious about the Development of Computational Humor. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico*. Morgan Kaufmann, 59–64.
- [35] Mengting Wan and Julian J. McAuley. 2016. Modeling Ambiguity, Subjectivity, and Diverging Viewpoints in Opinion Question Answering Systems. In *IEEE 16th International Conference on Data Mining, ICDM 2016*. IEEE Computer Society, 489–498.
- [36] Orion Weller and Kevin D. Seppi. 2019. Humor Detection: A Transformer Gets the Last Laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*. Association for Computational Linguistics, 3619–3623.
- [37] Janyce Wiebe and Rada Mihalcea. 2006. Word Sense and Subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, 1065–1072.
- [38] Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm Detection with Self-matching Networks and Low-rank Bilinear Pooling. In *The World Wide Web Conference (WWW ’19)*. ACM, New York, NY, USA, 2115–2124.
- [39] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard H. Hovy. 2015. Humor Recognition and Humor Anchor Extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal*. The Association for Computational Linguistics, 2367–2376.
- [40] Yasuhisa Yoshida, Tsutomu Hirao, Tomoharu Iwata, Masaaki Nagata, and Yuji Matsumoto. 2011. Transfer Learning for Multiple-Domain Sentiment Analysis - Identifying Domain Dependent/Independent Word Polarity. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA*. AAAI Press.