Sarcasm Detection on Facebook: A Supervised Learning Approach

Dipto Das
Anthony J. Clark
Missouri State University
Springfield, Missouri, USA
dipto175@live.missouristate.edu
anthonyclark@missouristate.edu

ACM.
ICMI '18,, October 16–20, 2018, Boulder, CO, USA
ACM 978-1-4503-5692-3/18/10.
https://doi.org/10.1145/3242969.3277524

Abstract
Sarcasm is a common feature of user interaction on social networking sites. Sarcasm differs with typical communication in alignment of literal meaning with intended meaning. Humans can recognize sarcasm from sufficient context information including from the various contents available on SNS. Existing literature mainly uses text data to detect sarcasm; though, a few recent studies propose to use image data. To date, no study has focused on user interaction pattern as a source of context information for detecting sarcasm. In this paper, we present a supervised machine learning based approach focusing on both contents of posts (e.g., text, image) and users' interaction on those posts on Facebook.

Author Keywords
Sarcasm; Sentiment; Text; Image; Facebook; Supervised Learning

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Introduction
Social networking sites (SNS) are a major medium of communication. People assess the sentiment of contents shared on SNS by considering all of the various aspects of those

Terminologies

Description: Posts shared on SNS may include a short explanation about the contents' gist, source, and audience.

Message: The user who posts content on SNS may choose to associate the content with a message written by him/her. He/She expresses what he/she thinks about the post, describes the content in detail if necessary, mentions various topics regarding the content like place, persons, feelings, etc.

Reactions: Almost all SNSs provide users with some ways to react to contents on those platforms. Some SNSs provide a like/star based system (e.g. Instagram, Twitter), and some provides upvote-downvote based system (e.g. Reddit, Quora). Since February 24, 2016, Facebook has supported a six-types-of-emotions (like, love, haha, wow, sad, angry) based react system.

contents together. A good amount of studies have focused on sentiment analysis on SNS. However, sarcasm is often hard to detect because people convey negative sentiments using seemingly positive words and vice-versa. Thus, it has not gained as much attention as straightforward positive-negative sentiment analysis. Most existing works depend only on text data for detecting sarcasm. A few recent works propose that multimedia contents (e.g., image) shared with the posts can be useful to detect sarcasm [3, 9]. Previous works have identified the importance of context information for detecting sarcasm [2]. However, to the best of our knowledge, no study has taken user interaction on a post as an indicator of detecting sarcasm. We propose that users' interaction on a post can be helpful to understand the context and thus can help to detect sarcasm. Here, by user interaction we mean the way they react to a post on SNS with reaction buttons or with comments. For testing our hypothesis, we needed multimodal SNS data. We take text, image, and user interaction into consideration. We developed a supervised learning model that can detect sarcasm on Facebook (FB) data with 93% accuracy. The major contributions of this work are: considering user interaction as an indicator of sarcasm, and a supervised learning model for detecting sarcasm.

Related Works

Tepperman et. al. [11] first worked to address the problem of detecting sarcasm on social media. They proposed experiments to recognize sarcasm using contextual, prosodic cues, but given the limited capability of NLP at that time, they took a naïve approach–detecting sarcasm based on use of the phrase "yeah, right". Later works on sarcasm detection by Filatova et. al. [4], and Bamman et. al. [2] emphasized on context for detecting sarcasm. In their works, Filatova et. al. [4] and Riloff et. al. [8] indicated that the contrast between positive and negative sentiment can be a sign of sarcasm in twitter. According to them, presence of positive and negative sentiment yielding words or phrases in a tweet can denote the that tweet to be sarcastic. The works by Das et.al. [3], and Schifanella et. al. [9] emphasized the importance of considering images besides text to detect sarcasm. Schifanella et. al. [9] suggests that the difference between sentiments yielded by the caption and the image is an indicator of sarcasm. According to Das et. al. [3], even without captions, images alone can convey sarcastic cues, and they proposed a CNN-based model for detecting sarcasm on Flickr. However, no works so far have focused on users' interaction around a particular content to detect sarcasm. In this study, we considered several aspects of SNS posts to detect sarcasm on a popular social networking site Facebook: the sentiment of posts, and the nature of user interaction around those posts. Users might chose to interact with text, images or use of reaction buttons.

Data

Previous works to detect sarcasm employed various methods for data collection. Some researchers used hashtags (e.g., #sarcasm) as indicator of sarcasm on twitter, some used context-based approach for identifying sarcastic contents on SNS [5]. We used the Facebook Graph API to collect data. Our data was collected after Facebook adopted the GDPR guidelines. For collecting SNS posts with sarcastic intents (i.e. positive instances) and posts with non-sarcastic intents, we selected ten public sarcasm related pages (e.g., Sarcasm Society) and verified FB pages of ten popular mainstream news media (e.g. The New York Times) on FB that have at least one million followers. We collected the description, message, image (if any), reactions, and comments (without users' identifying information). We only collected contents posted on FB with 'public' privacy setting. Since, reactions are a relatively new fea-

Figure 1: Sample of a Facebook post. (1) Message of the post; (2) Image of the post; (3) Description of the post; (4) Count of users' reactions to the post; (5) Users' comments on the post

tures, introduced on February 24, 2016, we chose to collect contents posted after February 2016. In total, we collected 20,120 posts of sarcasm category (48.65%) and 21,230 posts of non-sarcasm category (51.35%). Among the posts we collected, 98.26% posts include an image.

Experiment

We have three types of data: numeric, text, and image. Descriptions, messages, and comments are text data (1, 3, 5 in Figure 1). For most posts, there is an associated image (2 in Figure 1). Each post also has a count for six types of reactions (4 in Figure 1). The message, image or any other part of the post might not be sarcastic on its own but they altogether might convey sarcasm. In this study, we are trying to detect whether a post as a whole conveys sarcasm.

Reaction Data Pre-process

The six reaction counts on posts are the only numeric input data. They were first introduced on February 24, 2016. We considered the rest of that month as a burn-in period for the users to get familiar with these because many users might kept using the reaction buttons and that could harm the pattern of their usage in updated platform. Another concern is that, the reactions received on a post varies with how much reach (i.e. to how many users FB showed that post) a post receives on FB. Since the algorithm FB uses to arrange users' newsfeed is not known, we chose to use normalization. We divided the number of each reaction by the total number of received reactions on each post to remove the bias created by posts' reach.

Sentiment Analysis

For textual sentiment analysis, we considered two properties: subjectivity and polarity. Many existing works report context and sentiment as useful sources of sarcasm detection [4, 2, 8]. Subjectivity means the amount of expression of a user's sentiment, feelings, or opinion in a piece of text. Polarity denotes whether the text yields positive or negative sentiment. We used TextBlob [6] for determining subjectivity and polarity. Subjectivity is measured in scale of [0, 1] and polarity is measured in scale of [-1, 1]. Text with subjectivity near zero does not convey much information about a user's feelings (e.g. names that are tagged in text). A polarity value less less than zero indicates negative sentiment, while polarity greater than zero means that a user expresses positive sentiment with that piece of text.

Though a post can have more than one comment, it can have at most one description, message, and caption of image. For the latter three data, we determined the subjectivity and polarity of the text. Thus, we get six sentiment based features from textual data. However, since a post can have multiple comments, we modified the technique for sentiment analysis. For each comment, we calculated the subjectivity, positive sentiment (if polarity>0), and negative sentiment (if polarity<0). We used sum of subjectivity scores, sum of all positive sentiment scores, and sum of all negative sentiment scores of all comments as three individual features.

Image Caption Generation model

Schifanella et. al. [9] explored the importance of considering visual and textual aspects of SNS contents. They used semantic representations of the images. However, we argue that captions of images can provide semantic representations and hint about the sentiment expressed by an image at the same time and thus provides more useful information for detecting sarcasm. For automatically captioning images, we used the image captioning model proposed by Vinyals et. al. [12]. Each image now has a model-generated caption. Besides, it also might have a user-given caption.

Feature	*Information Gain*
Reaction [a]	
angry	0.3217
haha	0.4904
like	0.5534
love	0.4275
sad	0.3328
wow	0.4493
Image Data	
auto caption	
polarity	0.0174
subjectivity	0.0173
CNN score	0.0263
Text Data	
comments	
negativity	0.2503
positivity	0.4185
subjectivity	0.4626
description	
polarity	0.0237
subjectivity	0.0253
message	
polarity	0.1825
subjectivity	0.2044

Table 1: Information Gain of Features

[a]Specific to Facebook platform

Image Sarcasm Detection Model

We used a CNN-based model proposed by Das et. al. [3] that can detect sarcasm with 84% accuracy from images based on the visual cues. If an SNS post does not have a description or a message associated with it, the image is the only medium for knowing if the post has sarcastic intent. We pass the image of each post to this component and it outputs the probability (we call *CNN score*) of this image to have sarcastic cues in it.

Model Training

From the collected dataset, we constructed the 16 features listed in Table 1. We used scikit-learn [7] for machine learning algorithms. For missing values of any feature (e.g., caption subjectivity, caption polarity, *CNN score* if there is no image with a post), we used the average value of that feature as the representing value. We used 10-fold cross validation approach for validating our models. We used five supervised machine learning algorithms as follows: support vector machine (SVM) with linear kernel, two ensemble algorithms: Adaboost with Decision Tree classifier of depth 1, and Random Forest with scikit-learn's default parameter values, Multi Layer Perceptron (MLP), and Gaussian Naïve Bayes.

Result

Among all features, only reaction counts (like, love, haha, wow, sad, angry) are specific to FB. The other ten features are general to any SNS plaform. Table 1 shows the Entropy (a measurement of impurity) based information gain (reduction of entropy by using a particular feature) of our features. Information gain can be used to rank the features, higher information gain indicates that a feature will be more useful to machine learning algorithms [10, 1]. In Table 2, we present accuracy results for several different classifiers; stochastic algorithms were repeated 25 times.

Algorithm	SVM	Ada Boost	Random Forest	MLP	Gaussian NB
$Acc. \pm S.D.$	88.39 ± 0.0	90.61 ± 0.0	93.11 ± 0.196	92.06 ± 0.190	73.66 ± 0.0

Table 2: Applied ML Algorithms, Accuracies with Std. Deviation

In our study, we used a bag-of-features approach. Each feature we used can be used to build a weak classifier for sarcasm detection. Therefore, it was expected that ensemble approach combining these features will be a good classifier. We can see that both ensemble algorithms we used—Random Forest and AdaBoost performed very well for sarcasm detection. MLP-based model even with a small number of layers and nodes also performed well (>90% accuracy). SVM-based model's performance was not as good as ensemble models. Again, Naïve Bayes (NB) algorithms are widely used for text, sentiment data analysis. Since the features we are considering have continuous values, we chose to use Gaussian NB. It is surprising that Gaussian NB's performance was worse than that of other models.

Conclusion

In this extended abstract, we presented our findings for detecting sarcasm on social media using supervised learning algorithms on a noisy dataset (the data could include duplicate images and spam messages). Our results show that supervised learning algorithms, especially ensemble algorithms, are good fit for such applications. As part of our continuing work, we will study the impact of spam messages and duplicated data on the accuracy of sarcasm detection. Additionally, we are working to generalize our methods so that they work on other SNSs.

REFERENCES

1. Taqwa Ahmed Alhaj, Maheyzah Md Siraj, Anazida Zainal, Huwaida Tagelsir Elshoush, and Fatin Elhaj. 2016. Feature selection using information gain for improved structural-based alert correlation. *PloS one* 11, 11 (2016), e0166017.
2. David Bamman and Noah A Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *International AAAI Conference on Web and Social Media (ICWSM)*. 574–577.
3. Dipto Das and Anthony J Clark. 2018. Sarcasm Detection on Flickr Using a CNN. In *International Conference on Computing and Big Data (ICCBD)*.
4. Elena Filatova. 2012. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In *International Conference on Language Resources and Evaluation (LREC)*. Citeseer, 392–398.
5. Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*. Association for Computational Linguistics, 581–586.
6. Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, and others. 2014. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing* (2014).
7. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
8. Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 704–714.
9. Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1136–1145.
10. Bangsheng Sui. 2013. *Information gain feature selection based on feature interactions*. Ph.D. Dissertation.
11. Joseph Tepperman, David Traum, and Shrikanth Narayanan. 2006. " Yeah Right": Sarcasm Recognition for Spoken Dialogue Systems. In *Ninth International Conference on Spoken Language Processing*.
12. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.