Sarcasm Detection in News Headlines

Ceren Akkalyoncu 2657341
Tom Boersma 2509476
İrem Gökçek 2657359
Kevin Rojer 2520680

Abstract

Sarcasm is an ironic or satirical expression of ones opinion. However, identifying sarcasm is tough because of the distinction between the intended and the literal meaning. The goal of this paper is to decide whether a news headline is sarcastic or not using several natural language process approaches. The data originates from English language news headlines.

1 Introduction

This sentence introduces the best article you will ever read. Probably, you presumed that last statement might be sarcasm. According to the linguist Raymond Gibbs (1986), sarcasm includes "the use of words to express something other than and especially the opposite of the literal meaning of a sentence." (4) It is a form of irony that usually targets a specific individual.

However, it is very complicated to determine whether someone is sarcastic. Particularly, in the digital age that has transformed the way individuals communicate with each other. Texting, e-mailing, online commentaries and reviews are replacing face-to-face conversations. More and more people use sarcasm to add humor and cynicism to their style of communication. Sarcasm could lead to miss-communications caused by disambiguation. Therefore, researchers have been trying to detect sarcasm using the help of artificial intelligence.

In natural language processing (NLP), detecting sarcasm is a significant obstacle. Ravi and Ravi (2015) (1), in their detailed review of opinion analysis, mention that little or no studies have been devoted to study of irony or sarcasm detection, making it an open problem in the area. Their research shows that since there are no visible physical cues, judging whether a text is sarcastic or not is more laborious than deciding it in real life conversations. There are no facial cues, no vocal tones and perhaps the signals arrive with a delay because someone cannot reply immediately. Also, if someone does not know the person all that well, the last potential cue might get missed: history.

The automatic detection of sarcasm requires a system that can interpret the emotional language used in the text. Moreover, in many cases context is needed to understand what people are saying. This research focuses on news headlines. They are designed to be short and to attract attention. However, due to their grammatical structure, news headlines can lead to ambiguity. In contrast, entertainment news sources like The Onion use sarcasm in their headlines to satirize the nature and format of traditional news.

The kaggle competition located at `https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection/` gave us the inspiration to work on this subject. We will be trying sentiment analysis, opinion mining and other natural language techniques to decide if the news headlines are sarcastic or not.

2 Data Description

In our dataset, the instances are the news articles. An instance composed of the articles link, its headline and whether it is sarcastic or not. The dataset consists of 26709 articles from two websites: Huffington Post where the news headlines are factual and The Onion of which the news headlines are sarcastic. Previous studies were made using Twitter datasets where the data would be tweets. However, there would be too many grammar errors and confusion of concept. Also, considering that some tweets are just replies to another tweet, it would have been harder to label them. Since our data is news headlines, there are no grammar mistakes and the subject can be understood easily. Moreover, The Onion is less noisy compared to Twitter when finding sarcastic texts. The data in kaggle competition is in json file. All the articles taken from The Onion are sarcastic and all the rest that is not-sarcastic is taken from The Huffington. At first, we thought this could damage our results, but since we didn't choose the articles link as a feature, it wouldn't cause us any trouble.

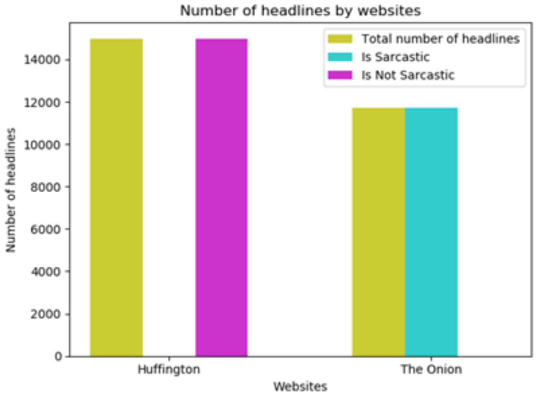

Figure 1: Sarcastic Non Sarcastic Headline Distribution per Website

3 Data Processing

Before starting with the modeling, we first processed the data to make it more meaningful for the models. We normalized the sentences by converting all letters to lower case. To make identifying of the part of speech tag of a given word possible, we lemmatized the headlines, using SpaCy and NLTK. We then removed the English stopwords, which did not provide any knowledge about the meaning of the sentence. At last, all the punctuations were removed. As an example, the first headline Former versace store clerk sues over secret 'black code' for minority shoppers, became versace store clerk sue secret black code minority shopper" after processing.
Cleaning the data helps with the understanding of the meaning of a given text, as it gets rid of all the unnecessary parts in text and we're only left with the parts that give us more information and insight about the context.
We also examined the structure of the data and checked the most common words in the processed dataset. As the dataset is based on news providers from US, there were a lot of context related words: Trump, Obama, American etc.

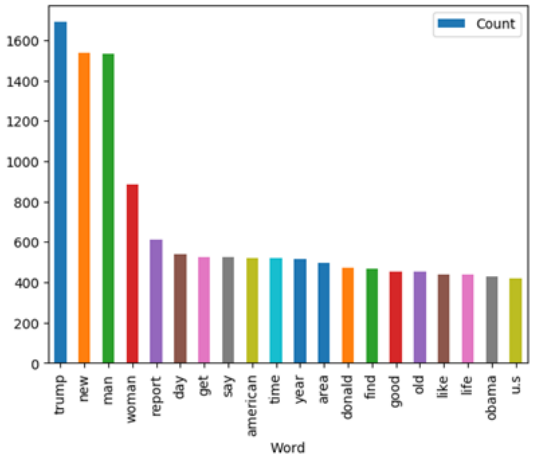

Figure 2: Most common words

4 Sentiment Analysis

Sentiment analysis is a natural language processing approach to identify the emotional tone of a text. This includes techniques like emotion detection or the determination of the level of polarity. News headlines rarely show forms of emotions, therefore the choice is made to only use the level of polarity. These levels are calculated for each level through the open-source package 'Vader', which assigns four different scores to each headline. A level of negativity, neutrality, positivity and compound. These levels are calculated according to polarity scores of each word in the headline.

The specific choice has been made to add all four scores seperately. As the contrast in polarity within headlines, through juxtapositions, could be a good indicator for sarcasm. Take for example a positive sentiment followed by a negative activity.

The detection of sentiment can be challenging. Sentences with a neutral sentiment are often misidentified and pose a problem for the model. The level of polarity can also be wrongly scored when sarcasm is not detected, hence falsely assigning a positive or negative score. Furthermore sentiment can be hard to identify because the system does not understand the context or tone. Context plays an important role in most sarcastic headlines.

5 Topic Modeling

Topic modeling is an NLP technique based on an unsupervised learning approach for discovering the topics that occur in a collection of documents. It stems from the idea that some latent variables which are not directly observable govern the semantics of the documents. As a result, topic modeling uncovers, recognize and extract these latent variables - topics - covered in the document.

The concept of topic modeling is crystal clear. The news headlines are clustered together where each word in a headline will be processed to discover the topic. Each headline will be assigned a value based on the distribution of these words.

In particular, Latent Dirichlet Allocation (LDA) is the foundation of the topic modeling technique used in this research. Conceptually, the LDA is a probability distribution over distributions. It is a multivariate probability distribution that describes the distribution of categorical distributions (Blei et al., 2003). In this case, the categories are the topics identified by the system. The output of the LDA model is a vector representing its topic distribution for a given document. For example, 5% topic 1, 48% topic 2, 10% topic 3, 2.7% topic 4, etc.

Human-interpretable topics are extracted from the news headlines using the Python library genesim and pyldavis. First, a dictionary containing the words from the data is created, and a bag-of-words corpus gets formed. Then, the predefined number of topics is chosen and used as an input parameter for the LDA model. In this research, LDA was asked to find 20 topics. For each topic, the words they are most strongly associated with characterize the topic.

Using the library pyldavis, users can interpret the topics in a topic model fitted to a corpus of text data. A user can extract information from a fitted LDA topic model to create beautiful visualizations.

6 Results

The first model achieves a reasonable performance as can be seen in the contingency table in Figure 3. The model has an overall error rate of 0.228, with an average precision and recall of 77% and 76%. Noteworthy is the difference in recall between the sarcastic and non-sarcastic headlines, which achieves high scores for non-sarcastic headlines of 85%. However the recall for sarcastic headlines scores a lot lower, namely 67%, hence misclassifies a lot more sarcastic headlines in comparison to non-sarcastic headlines.

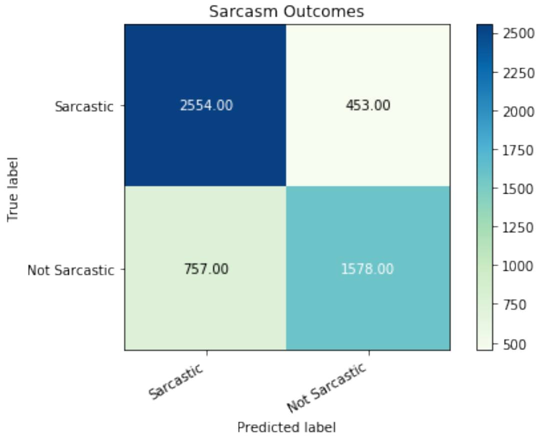

Figure 3: Confusion Matrix

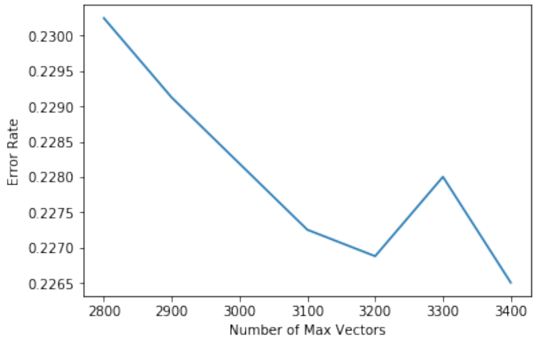

Figure 4: Error Rate

The addition of the four levels of polarity decrease the error rate of the model, but only by 0.5%. It slightly improves the recall of the sarcastic headlines, by 1%. The 'Vader'-package has a hard time deciphering the sentiment of the headlines, as can be seen in Figure 5. There is little to no difference between the four polarity scores of the sarcastic newspaper: 'The Onion' vs the non-sarcastic newspaper:

'The Huffington Post'.

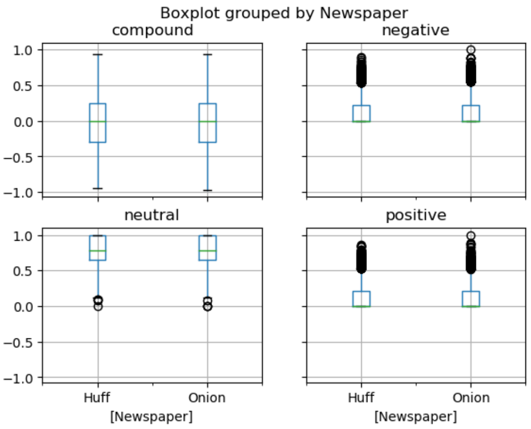

Figure 5: Polarity scores

Using LDA to fit the dataset, we explored the topic model to interpret the meaning of the different topics. However, this is a difficult task because the output of the model is a big probability mass function over all possible words in the model for each individual topic. So, we could represent the each topic by their frequency of the words associated with it. For example, we could visualize topic number six by ranking its 30 most frequent words associated with the topic in Figure 6. As one can infer, the topic probably has to do with The Government of the United States.

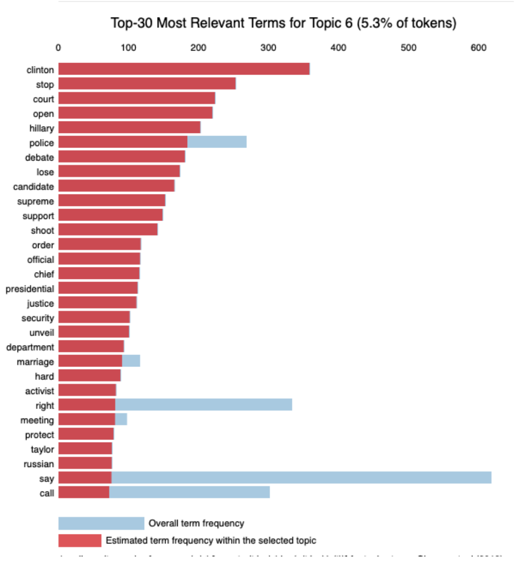

Figure 6: The 30 most strongly associated words of topic number 6.

Furthermore, in Figure 7, one can look at the underlying structure of the topic. The size of the bubbles indicate the importance of the topics, relative to the data. Bubbles that are close to each other should be semantically related. However, after closely inspecting the 30 most relevant words for the neighbouring clusters no real connections could have been identified. Therefore, the conclusion is that topic modeling for small bodies of text such as news headlines, do not work effectively. The reason for this could be the fact that due to their shortness and grammatical structure, news headlines lack enough context to accurately classify the topics. Another possibility could be the number of topics chosen for the LDA model. However, adjusting for the parameter value didn't yield any clear associations between the topics.

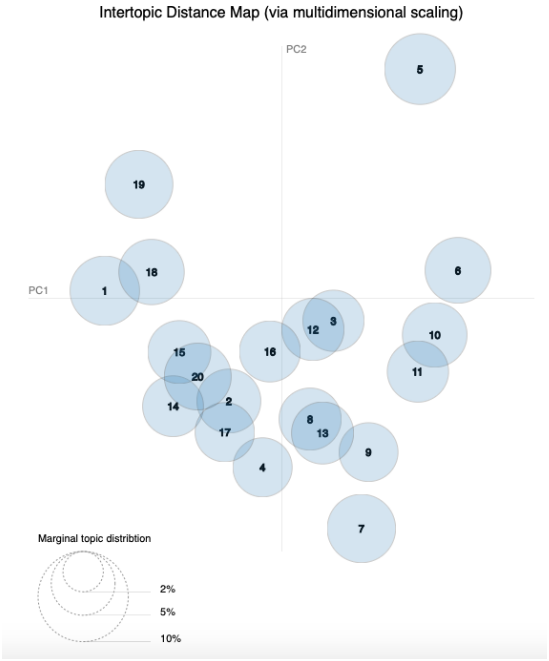

Figure 7: Topic cluster

7 Conclusion and Discussion

Sarcasm is a common occurence, mainly in spoken language. Lexion-based text mining techniques fail to make distinction between sarcastic and non-sarcastic sentences as sarcasm doesn't have a comprehensible pattern and is highly dependent on the context.

In this project, we worked on detecting sarcasm, aiming for a reasonable accuracy that could be trusted in usage. By extracting information from the sentences that could be used for detecting sarcasm, we managed to identify it with considerable success. However, as our dataset only consisted of news headlines in professional language, this model would not be applicable to a dataset of another context, like identifying sarcasm in tweets which require further processing and also more thorough feature extraction to get a good result.

As sarcasm is context dependent, having a general knowledge of the given context could be considered for further research. Also, for more publicly written platforms such as Twitter, a research on a personal level - inspecting the writing style or past tweets might give interesting results for a more thorough opinion mining and sarcasm detection. While these may be expensive to manage for varied topics, it may be worthy to experiment on specific contexts.

As discussed in the lectures, majority of the information is stored in texts and extracting knowledge from text is an important problem. One aspect of information extraction is sentiment analysis and opinion mining (2). Using the existent techniques for opinion mining, we can only derive basic meanings. These kinds of approaches often ignore the complexity of the natural languages. This often includes sarcasm, a complex usage of language that implies mockery or ridicule in a harsh way.

Our purpose was to identify sarcasm in news headlines. In this project, we divided the work as follows:

İrem was in charge of data inspection, data processing and applying logistic regression model. Kevin was in charge of topic modelling and data processing. Tom was in charge of applying sentiment analysis. Ceren was in charge of data inspection, data processing and applying multinomial naive bayes. Moreover, we helped each other and worked together in every process. Our project can be found at this repo: https://github.com/gokcekirem/SarcasmDetector

References

[1] Ravi, K. and Ravi, V. (2015), A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowledge-Based Systems, Vol. 89, pp. 14-46.

[2] Yee Liau, B. and Pei Tan, P. (2014), Gaining customer knowledge in low cost airlines through text mining, Industrial Management Data Systems, Vol. 114 No. 9, pp. 1344-1359.

[3] Shubhadeep Mukherjee, Pradip Kumar Bala, (2017) "Detecting sarcasm in customer tweets: an NLP based approach", Industrial Management Data Systems, Vol. 117 Issue: 6, pp.1109-1126, https://doi.org/10.1108/IMDS-06-2016-0207

[4] Gibbs, Raymond W. (1986) "On the psycholinguistics of sarcasm.", Journal of Experimental Psychology: General, Vol. 115 (1), Mar 1986, pg. 3-15

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, (2003), "Latent Dirichlet Allocation", Journal of Machine Learning Research 3 (2003) pg. 993-1022