International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 04 – Issue 08 || August 2019 || PP. 104-108

A Comparative Analysis of Sarcasm Detection

Atul Kumar
Research Scholar DSMNRU, Lucknow

Ratnesh Kumar
Computer Science Dept SRMGPC, Lucknow

Dr. Vinodini Katiyar
Professor DSMNRU, Lucknow

Abstract: The sentiment is an attitude or opinion that is expressed, it is not based on laws and philosophy. Sentiment analysis is the mining of opinion or sentiment expressed in text as positive, negative or neutral sentiment. Detecting sarcasm is the main hurdle in sentiment analysis. Sarcasm can be expressed in various forms like in conversation, heading or title of the novel. As Sarcasm represents contrary sentiment to the literal meaning that is conveyed in the text, it is hard to identify sarcasm even for a human. This paper presents a study on sentiment analysis. The datasets, feature engineering, and algorithm used in previous models for sarcasm detection. The study proves that Support vector machine (SVM) is the finest approach for sarcasm detection, feature engineering is an import assent in training a model and as News Headline dataset is written by professionals, it is the most suitable data set to work on.

Index Terms: Sentiment Analysis, SVM, Sarcasm.

I. INTRODUCTION

Language is an important medium for human communication. It allows us to convey information, express our ideas, and give instructions to others. According to some philosophers, it enables us to form complex thoughts and reason about them. What makes it so hard for computers to understand us? One feature of human language, or a drawback depending on how you look at it, is the lack of precisely defined structure. Structured languages are easy to parse and understand for computers because of the strict set of rules or grammar. The Language we used to communicate with each other also has defined grammatical rules. But mostly, human discourse is complex and unstructured. So, what can computers do to make sense of unstructured text? Computers can perform a level of processing with words and phrases, identify keywords, parts of speech, named entities, dates, quantities, etc. using these they can parse some structured sentence and extract relevant parts of statements, at the higher level.

To understand the proper meaning or semantics of the sentence we implicitly apply our knowledge about the surroundings. For example, in TABLE 1 we know that wide things don't fit through narrow things. There are many other scenarios in which some context is indispensable for understanding what is being said. A system should understand the relationship of the words, the context surrounding the utterance. Sentiment analysis is a process containing methods, techniques, and tools to detect and extract authors' attitude toward document and subjective information, such as opinion and attitudes like positive, negative and neutral, from the language. In sentiment analysis, Sarcasm plays a key role in accuracy. As Sarcasm represent opposite sentiment to sentence. Detecting sarcasm and in sentiment analysis will boost accuracy. As most of the modern industry depend on the sentiment of reviews to rate the product and use it in regular activity for stronger growth of the industry.

Sarcasm detection is one of the difficult problems in sentiment analysis. Sarcasm is an insincere statement designed to provoke or insult. This commonly serves as a means to insult or humour in a sentence. In most cases the intended meaning is different from general meaning. Sarcasm can be expressed in several ways for example in a conversation or text or heading in newspaper or it can be a title of a web series/ novel. Expressing Sarcasm in conversation mainly depends on tune of speaker, the facial expression of the speaker, body language and the situational environment in it's said. For example

- "Sometimes I need what only you can provide: your absence." – Oscar Wilde
- "I never forget a face, but in your case I'll be glad to make an exception."-Groucho Marx
- "I worked my whole life, to be this poor. YAY !!"

Detecting Sarcasm in this case can be done by tone analysis i.e. Change in tone. But, Detecting Sarcasm in sentence whether it is tweet or heading is most difficult. Here Sarcasm can be expressed in

Capitalized words, use of emojis, use of positive and negative words, etc. Over the years of research, different approaches were created, like rule-based, statistical, and example-based Sarcasm detection. With all this effort, it's still an unsolved problem. However, few methods have made a large leap forward in Sarcasm detection.

II. Sarcasm study

Sarcasm is derived from the French word 'sarcasm or', and also from the Greek word 'sarkazein', which means "tear flesh," or "grind the teeth". Social media such as Twitter exhibit rich sarcasm phenomena and recent work on automatic sarcasm detection has focused. Detecting sarcasm automatically is useful for opinion mining and reputation management and hence has received growing interest from the natural language processing community. Sarcasm detection is a classification problem i.e. a text/ line is sarcastic or not. Sarcasm is closely related to irony in fact, it is irony.

A. Characteristics of Sarcasm

According to different authors, sarcasm arises when there is irony between text and contextual information. For example, in the sentence "I love working in holidays!" as marked as sarcastic because of irony in the sentence, as working in holidays is an undesirable condition and the speaker is saying to love that condition. As per author sarcasm can be identified by analyzing the response to that sentence. The response can be a burst of laughter, smile, change of topic, silence, and sarcasm in return.

B. Types of Sarcasm

Sarcasm arises when there is chaos between literal meaning and the intentional meaning of the text. Sarcasm can be divided into seven classes, namely Coexistence between positive and negative negative sentence followed by a positive sentence and vice versa A dilemma in the sentence Negative phase followed by a positive phase and vise-verse Comparison between worse and worst Comparison with something better Incongruity in sentence No specific positive and negative point Sarcasm is dependent of situation and surrounding of the speaker =or where the conversion is held. Sarcasm is hard to detect, also for a human. It can be represented in many ways.

III. Data

For training an appropriate model to classify that a text is sarcastic or not, depends on dataset it is developed on. Dataset is the initial step of training a model. The most frequently used dataset for training models is a Twitter dataset, Amazon product review and News headline dataset. The datasets consist of a reasonably large amount of text data that are sarcastic and non-sarcastic. They all are messy and consists of many irregulars and missing entries. To develop a model on these datasets, they need to be cleaned and preprocessed first.

A. Twitter Dataset

Twitter dataset is a massive collection of daily tweets from users. It can be obtained using API and tools like tweet4j. the tweets are text data consists of a username, URL, hash tag, and text. For example, in the tweet "Congrats bro @user, Keep working you will do it in eons. https://twet.ts/goforit#sarcastic ", "@user" is a username, "https://twet.ts/goforit" is a link and "#sarcastic" is a hash tag. There are a few tweets that consist of meta tags. Tweets can be nested i.e. contains tweets in tweets which makes it harder to model to gain from it. There are two types of tweets that are mainly required, one is sarcastic, and another is normal tweets. To collect Sarcastic tweets, tweets with the sarcastic hashtag (#Sarcasm) is collected. As it contains data that are posted by the general community, there are spelling mistakes and informal usage. The precisive sarcastic data is bad.

B. Amazon Product Review Dataset

Amazon product review dataset is a collection of review of a product consisting of stars, title, date, an author, product, and review. The reviews are formatted in the mark-up language. The corpus contains reviews from Amazon websites and is classified as ironic and regular. We can consider ironic reviews as sarcastic to train our model. As the data is formatted in the mark-up language, tags should be removed. As corpus contains irony as sarcasm the precisive sarcastic data is higher than twitter data. The irony can be in text or different polarity of review and stars, of a product in the corpus

C. Headline Dataset

This dataset was collected from The Onion and Huff Post. It contains sarcastic and regular news

headlines. The sarcastic news headlines were gathered from The Onion and normal headlines from Huff Post. As it is not written by the general population, the chances of spelling mistakes and informal usage are low. It contains 27K of headlines, as of it 11.7K are sarcastic and 14.9K are non-sarcastic. The dataset consists of three attributes, is Sarcastic, Headline and link. The is Sarcastic points out whether the instance of headline is sarcastic or not, Headline attribute contains the headline of the article and link attribute contains the line of the news article. The precisive sarcastic data is greater.

IV. Feature Engineering

The most significant part is feature engineering, it's not using a complicated model but analyzing and extracting which features to extract, and how they are related to giving a better model. Grammar plays an import role in feature extraction. Features can be classified into four categories,

- Lexical Features
- Pragmatic Features
- Explicit Incongruity
- Implicit Incongruity

A. Lexical features

Sentences containing specific words increase the chances of the sentence being sarcastic like "Yeah!! As if". Few terms that are highly contiguous for sarcastic and are taken special account. N-grams is the most frequently used Natural language processing tasks. Each word is given a unique number and a dictionary is created for a feature vector. TF-IDF is another way to create it as most of the items will be zero in the vector.

B. Pragmatic Features

Pragmatic feature deals with grammatical issues. It deals with capitalization, emotion, laughter expression and Punctuation marks. In a certain situation, we use capitalization to make a point. In the same way, sometimes sarcasm is also capitalized to make a point. Excess use emotion and laughter expression can be an indication of the sentence being sarcastic. Since sarcasm is mocking with humor, excess use words like "lol", "bingo" can be an indicator of sarcasm. Punctuation marks are used as an extra base of emotion like surprised and amazed e.g."!!".

C. Explicit Incongruity:

Explicit incongruity is expressed with direct sentiment keywords. It contains sentiment words of both polarity i.e. positive and negative, in a text and vice versa. Sarcasm mostly contains a sentence with positive sentiment followed by negative/ bitter sentiment, and vice versa. For example, "I love being Ignored" here "love" and "ignored" are opposite polarity sentiments.

D. Implicit Incongruity

Implicit incongruity is expressed in a deeper meaning or indirect Sentiment phrases. It contains words with positive sentiment words followed by an indirect negative sentiment and vice versa in a text. For example "Your singing is amazing, its´ like Albert Einstein is singing in front of me." Here "amazing" is a positive sentiment whereas "Albert Einstein is singing in front of me" is a negative sentiment as Albert Einstein doesntsing.

V. Methods

A model is a machine learning algorithm based on features to predict whether a text is sarcastic or not. More there are many approaches to solve the sarcasm detection problem. It is a classification problem i.e. classifying whether the text is sarcasmtic or not, it can be solved using the following ways.

A. Support Vector Machine

Support vector machine mainly deals with margins and boundaries. Author [2] used 2 baseline models one with unigram features and other with unigram bi-gram and tri-gram features. Author [4] used one class SVM model to classify texts.

B. Naïve Bayes

This algorithm is based on probability. It is mostly used in the classification problem. Author [3] used and found it doesn't predict the best result in this case.

C. Random Forest

This machine learning algorithm is efficient when having a fewer data instance but large features columns. Author [6] used a random forest with the 500 number of trees and 3 variable.

D. Lexical Method

It is a rule-based method where some specific words mean the presence of Sarcasm. Author [] used the lexical approach to solve the problem.

E. Neural Networks

It is a new approach to train the neural network model. It is deep learning based. Author [5, 6] used this approach L2 parameter 0.001and learning rate of 0.01.

Table **??** shows different approaches used by authors with their F score and dataset used.

Table I
Approaches Used by Different Authors

AUTHOR NAME	APPROCH USED	Dataset	Accuracy(Fsc
[2] Piyoros Tungthamthiti	SVM (uni-gram)	Twitter dataset	0.76
	SVM (uni-gram, bi-gram,tri-gram)		0.79
[4] Chun-Che Peng	One class SVM		0.56
[5] Sahil Jain	SVM	Amazon	0.81
	Neural network		0.80
	Naïve Bayes		0.75
[6] Ashwin Bhat	Random forest	Twitter	0.59
	SVM(uni)		0.81
	Neural network		0.78

VI. Findings

From table 1 it can be concluded that support vector machine (SVM) is the best approach to construct a model to classify text as Sarcasm. the essential step to train a better model, feature engineering should be done and important features should be extracted like explicit and implicit Incongruity. Features play a key role in training a model. The Support vector machine (SVM) trained on both datasets have the same accuracy with both Amazon product review and twitter dataset. The SVM algorithm with unigram features includes works best with either of the datasets.

VII. Conclusion

Sentiment analysis is the mining of emotions associated with the text as positive, negative and neutral. The main obstacle in sentiment analysis is the presence of Sarcasm in the text. Sarcasm is a particular form of text which has different sentiment to the true sentiment in the text. The News Headline dataset which is written by professionals consists of more precise sarcasm data as related to another dataset. It is written by professionals and contains no spelling mistakes and informal text. Feature extraction is a powerful step in creating an accurate model. Extracting feature from the dataset is important as better the feature, more accurate the model will be. There are mainly four types of feature extraction lexical, pragmatic, implicit and explicit Incongruity. Support vector machine (SVM) is the most accurate approach to solve the problem as compared to naïve Bayes, maximum entropy, etc.

VIII. References

[1] G, Dr. Vadivu Chandra Sekharan, Sindhu. (2018). A COMPREHENSIVE STUDY ON SARCASM DETECTION TECHNIQUES IN SENTIMENTANALYSIS.
[2] Piyoros Tungthamthiti, Kiyoaki Shirai Masnizah Mohd. (2018). RECOGNITION OF SARCASMIN TWEETS BASED ON CONCEPT LEVEL SENTIMENT ANALYSIS AND SUPERVISED LEARNING APPROACHES
[3] Christine Liebrecht, Florian Kunneman Antalvanden Bosch. THE PERFECT SOLUTION FOR

DETECTING SARCASM IN TWEETS NOT
[4] Chun-Che Peng, Mohammad Lakis Wei Pan. (2015) DETECTING SARCASM IN TEXT: AN OBVIOUS SOLUTION TO A TRIVIAL PROBLEM
[5] Sahil Jain, Ashish Ranjan Dipali Baviskar. (2018).SAR- CASM DETECTION IN AMAZON PRODUCT REVIEWS
[6] Ashwin Bhat, Yash Bhalgat, Kalpesh Patil Navjot Singh. (2017) SARCASM DETECTION INTWEETS.
[7] Elena Filatova. IRONY AND SARCASM: CORPUS GENERATION AND ANALYSIS USING CROWD SOURCING.
[8] J. M. Soler, F. Cuartero, and M. Roblizo, (2012) TWITTER AS A TOOL FOR PREDICTING ELECTIONS RESULTS.
[9] D. Maynard and M. A. Greenwood, (2014) WHO CARES ABOUT SARCASTIC TWEETS? INVESTIGATING THE IMPACT OF SARCASM ON SENTIMENT ANALYSIS.
[10] S. Homoceanu, M. Loster, C. Lo, and W.-T. Balke, (2011). WILL I LIKE IT? PROVIDING PRODUCT OVERVIEWS BASED ON OPINION EXCERPTS.
[11] Justin Martineau, and Tim Finin, (2009) AN IM- PROVED FEATURE SPACE FOR SENTIMENT ANALYSIS.
[12] Lakshya Kumar, Arpan Somani, Pushpak Bhat- tacharyya. (2017 DETECTING SARCASM IN NUMERICAL PORTIONS OFTEXT.