International Journal of Research in Engineering, Science and Management
Volume-1, Issue-11, November-2018
www.ijresm.com | ISSN (Online): 2581-5792

Sarcasm Detection in Social Media Posts: A Division of Sentiment Analysis

Mansi Vats[1], Yashvardhan Soni[2]
[1]*M. Tech. Student, Dept. of Information Technology, Dronacharya College of Engineering, Gurugram, India*
[2]*Professor and HoD, Dept. of Information Technology, Dronacharya College of Engineering, Gurugram, India*

***Abstract*: This paper is written before doing any practical work on the mentioned topic. The paper is based on the re-search done to find methods to detect sarcasm in the social media posts. The language that will be used for this is the R language. We have given a brief introduction of Sentiment Analysis, which sarcasm detection is a part of. We have defined and described Natural Language Processing (NLP), through which the machine can understand the human language to detect sarcasm. We have also described the steps for sarcasm detection and a comparison between different machine learning algorithms to choose for sarcasm detection. In the end, the result will be evaluated that whether the machine is able to detect sarcasm with atleast 80% accuracy or not.**

***Keywords*: sarcasm detection, social media, sentiment analysis**

1. Introduction

According to Merriam-Webster, sarcasm is defined as: "A sharp, and often satirical or ironic utterance designed to cut or give pain" OR "a mode of satirical wit depending for its effect on bitter, caustic, and often ironic language that is usually directed against an individual." Sarcasm is generally a statement which actually means the opposite of its obvious meaning. It is usually used to insult someone. We humans can identify the sarcasm in someone's speech by its tone. But when a user posts a sarcastic post on a social media platform, it becomes difficult for the machine to differentiate between the obvious meaning and the intended meaning of the post. Hence, it becomes important to detect sarcasm in social media posts. Sarcasm detection is still in its initial stages and is a complex task because there is no specific vocabulary for sarcasm, unlike spams. The same words can be used for normal usage as well as to express sarcasm. So, to detect sarcasm the machine would have to learn that how to figure out that the user means the opposite of his statement.

2. Sentiment analysis

Sentiment Analysis, also known as Opinion Mining and Emotion AI, is used to determine whether a statement is positive, negative or neutral. The main objective of sentiment analysis is to determine the attitude of the speaker or writer towards a certain topic or product. For this, we make use of Natural Language Processing, text analysis, computational linguistics, and biometrics.

Sentiment analysis is of great use in social media monitoring which helps us in obtaining the views or opinions of a wide range of public on particular topics. For example, if we want to know how the people of India find Chinese food, we can do this through sentiment analysis on social media. By analysing tweets of people on twitter, we can determine whether and why people find Chinese food good or bad. We can also make use of some exact words such as "very salty" or "very spicy" in order to have a better knowledge of why the consumers are not happy.

For social media monitoring, we can use a tool named Brandwatch Analytics to obtain the results in a faster and easier manner. Sarcasm detection is a part of sentiment analysis through which we can identify the sarcastic attitude of the speaker or the writer.

3. R Language

R is a programing language and environment used for statistical analysis, graphics representation and reporting. Statistical analysis include- linear and non-linear modelling, classical statistical tests, time series analysis, classification, clustering etc. R language is highly extensible. It was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. It is presently developed by R Developmeent core team. This language is freely available under GNU General Public License. It also gives pre-compiled binary versions for many operating systems like Windows, Mac, Linux etc. Its name was given as R on the basis of the first letter of the first name of the two of its authors. The core of R is an interpreted computer language. It permits branching and looping and also modular programming using functions. It also permits concatenation with the procedures written in C, C++, Python, .Net or FORTRAN languages for productivity.

A. Features of R language

It is a well-developed, simple and potent programming language which comprises of conditional loops, user-defined recursive functions, and input and output facilities.

- It provides effectual data handling and storage space.
- It provides a package of operators for computations on arrays, lists, vectors, and matrices.
- It gives a huge, logical and non-segregated collection of tools for data analysis.

- It provides graphical provisions for data analysis and display either straight at the computer or printing at the paper.

4. Sarcasm detection

A. *What is sarcasm?*

From different sources, we get different definitions of sarcasm. Some of them are:

According to Cambridge English University- "The use of remarks that clearly means the opposite of that they say, made in order to hurt someone's feelings or to criticize something in a humorous way. For example, 'You have been working hard', he said with heavy sarcasm, as he looked at the empty page."

According to Collins Dictionary- "Sarcasm is speech or writing which actually means the opposite of what it seems to say. Sarcasm is usually intended to mark or insult someone. For example, 'What a pity', Graham said with a hint of sarcasm."

As a conclusion, sarcasm can be defined as a statement which means the opposite of its actual meaning and is generally used to affront, tease, taunt or censure someone. The sarcasm is generally taken in a negative sense, hence a sarcastic statement is a negative statement.

B. *Sarcasm in social media*

Social media platforms nowadays are full of a variety of posts. Many people are active on social media. People update each and every moment of their lives on social media sites such as facebook, twitter, instagram etc. Among these posts, some are sarcastic while some are non-sarcastic.

While posting sarcastic tweets or posts on social media, some people make use of hashtags such as #sarcasm for making it easy for the public to identify it as a sarcastic post.

While some people simply write sarcastic statements without hashtags, tag their friends there or reply to some tweets or posts in a sarcastic way.

Sarcasm can be expressed in many ways by, for example, reviewing a movie, match or commenting about the weather etc.

C. *Why to detect sarcasm?*

The number of social media users is increasing day by day. And a majority of users post a variety of posts on social media. This has led to a tremendous increase in the sarcastic posts on social media. The main reason behind this is that people find it sarcasm as more witty, humorous and an interesting way to express their feelings and criticize or insult others. Just like spams, sarcastic posts should also be identified individually. Sarcasm detection will also be helpful for the people who are suffering from Autism and Asperger's and who find it difficult to understand sarcasm, sardonicism and wit.

D. *Why is it difficult to detect sarcasm?*

Many data scientists are still working on sarcasm detection. It is a very complex task. This is because it becomes difficult for the machine to differentiate between the obvious meaning and the intended meaning of the post.it happens because of the same types of words used both in positive and negative sense as there is no specific vocabulary defined for sarcastic posts yet. The machine could not identify whether the statement is positive or negative.

To let the machine understand the human language more efficiently, we make use of NLP (Natural Language Processing).

5. Natural language processing

Natural Language Processing, abbreviated as NLP, is an AI method which is used for interacting with the intelligent systems using any natural language, for example, English. NLP can also be defined as automatic manipulation of natural language, which can be speech or text, by the program. Natural language is the manner in which humans interact with each other, which is speech and text.

Now-a-days, we see text in a variety of ways such as Signs, Menus, Email, SMS, and Web Pages etc. NLP is needed in a plenty of situations such as when we want a robot to act as per our commands or when we want help from a communicating expert system etc. the input and output can be provided to and from the NLP system in two ways,

- Written Text
- Speech

A. *Components of NLP*

There are mainly two components of NLP:

1) *NLU*

NLU stands for Natural Language Understanding. It depicts the given input in natural language into useful delineations. It also examines different characteristics of the language.

2) *NLG*

NLG stands for Natural Language Generation. It is a procedure for constructing significant phrases and sentences in the form of natural language from some inner portrayal. It involves

- Text Planning: It means recovering the useful data from the knowledge base.
- Sentence Planning: It refers to selecting the necessary words, constructing relevant phrases, setting quality of the sentence.
- Text Realization: It refers to plotting sentence plan into sentence structure.

The NLU is more difficult than NLG.

6. Steps in sarcasm detection

A. *Getting the data*

The first step in sarcasm detection is to collect data i.e. the posts, from the social media platforms. We should have wide range of source of collection and hence the data should be collected from various social media platforms like facebook,

twitter, instagram etc. collecting data from various sources will also help us in analysing the various levels of sarcasm on different social media sites.

Once the data is collected, the machine is trained through supervised learning.

Classification is a supervised learning method which is done to label some sentences as sarcastic while some as non-sarcastic ones in order to make our classifier.

B. Pre-processing the data

Pre-processing the data refers to cleaning up the data. After collecting the data from various sources, it is important to clean the data to remove the possibilities of having the posts in which sarcasm is either in the enclosed link or in the replies to other posts. To make sure that only the posts written in English are collected. The posts which contain Non-ASCII characters are removed. All the hashtags, friend tags and the use of word sarcasm/sarcastic is also removed from all the posts. Also the duplicate posts are removed.

1) Why data pre-processing?

Data pre-processing is referred to as a data mining technique which includes reconstructing the raw data into an interpretable form. The data prevent in the real-world is mostly incomplete, incompatible, lack some behaviours and has many errors. Data pre-processing is done to resolve such issues. It is used in database-driven environments such as Neural Networks.

C. Feature engineering

After the pre-processing i.e. cleaning of the data is done, next step is to perform feature engineering. Feature engineering is performed to find out the answers of the questions like:

What are the variables in a post that make it sarcastic or non-sarcastic?

How do we extract them from the post?

Feature engineering refers to the procedure of making use of the domain knowledge of the data to establish characteristics that make machine learning algorithms work. If feature engineering is performed in the right manner, it extends the prophetic power of machine learning algorithms by generating characteristics from unanalysed data that makes the machine learning process easier.

1) Importance of feature engineering

The characteristics in our data will directly affect the prophetic models we make use of and the outcome we can attain.

If we develop and select better features, we will attain better outputs. It is both true and deceptive.

The outcomes we attain are a part of the model we select, the data that is accessible to us and the characteristics we develop.

Great characteristics are required that report the construction innate our data. Better characteristics enable flexibility.

Even the selection of "wrong models" (less optimum) can lead us to good outcomes.

Better characteristics lead to simpler models.

D. Choosing a classifier

After feature engineering, the next step in sarcasm detection is to choose a classifier among the various machine learning algorithms. This is done to train the machine and help it classify the posts into sarcastic and non-sarcastic posts from the social media platforms.

This is the most crucial step and the selection of correct classifier is very important.

Classification is a supervised learning technique. By supervised learning we mean that the machine is trained by providing a set of input data and a teacher or "data set" is present to guide the machine to provide the correct results. Later, this learning is used to classify new statements.

The data set provided to the machine can be of two types:

- *Bi-class:* It helps in classifying the data into two categories. For example, whether a post is sarcastic or non-sarcastic.
- *Multi-class:* It helps in classifying the data in various categories.

Below are some of the classification algorithms in machine learning.

1) Naïve-Bayes Classifier (Generative learning algorithm)

Naïve-Bayes classification technique is based on the Bayes' Theorem. According to Naïve-Bayes classification, one feature of a class is not related to any other feature of the same class. Even if it happens that one feature depends on any other feature of the class, all of these attributes contribute individually to the possibility. This type of model is easy to use and is mainly used for huge data sets. It is a very simple algorithm and it also performs better than many highly advanced classification techniques. Equation for Naïve-Bayes algorithm is as follows:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

Where, P(c|x)=posterior probability, P(x|c)=likelihood, P(c)=class prior probability and P(x)=predictor prior probability.

2) Logistic Regression (Predictive learning Model)

It is a mathematical model to analyze a data set consisting of two or more independent variables that produce an output. The output is determined with the help of a dichotomous variable, which means a variable in which only two outcomes are possible. The main objective of logistic regression is to identify finest fitting model to narrate the association between the dichotomous feature of interest and a set of independent variables.

3) Decision trees

Decision trees are used to construct classification models in the structure of a tree. It converts the data set into smaller

fragments called subsets and a related decision tree is gradually developed simultaneously. The ultimate output is a tree with decision nodes and leaf nodes. A decision tree consists of two or more branches and a leaf node means the classification or decision. The highest decision node in a tree represents the best predictor and is called the root node. Decision tress can handle categorical as well as numerical data.

4) Neural network

A neural network is made up of small units known as neurons. The neurons are arranged in layers hence it is a layered architecture. These neurons convert the given input into output. The input is fed to each and every neuron of the first layer in the neural network. The neurons apply some functions on these inputs and then produce the output and pass it to the next layer. Usually, a neural network is feed-forward in nature i.e. the output produced by all the nodes of one layer are fed as the input to all the nodes of the next layer. But no feedback is facilitated. Every neuron in a neural network has a weight associated with it. These weights play a crucial role during machine learning and can be manipulated according to the desired output.

5) Nearest neighbour

The k-nearest neighbour algorithm is a supervised classification algorithm. By supervised we mean that it takes a cluster of tagged points and uses them to learn how to tag other points. To tag a new point, it examines the tagged points, adjacent to that new point, which are its nearest neighbours. It asks those neighbours to vote and the tag that has the highest number of votes is the tag for the new point. Here "k" is the number of neighbours it inspects.

E. Results and Insights

After performing all the above steps, the final step is to check whether the machine is able to detect sarcasm or not. This will be done by checking the output produced by the machine after training it. The produced or actual output will be compared with the desired output. If both the outputs match, it means the system is working fine. Else, the system will be trained again accordingly.

The sarcastic posts will be named as negative sentences and the non-sarcastic posts will be named as positive sentences. If the machine classifies the sentences correctly, it means that the training is successful.

7. Conclusion

This paper had concluded that sarcasm detection is possible by training the machine accordingly. Differ-ent data scientists may have different approach to-wards sarcasm detection. Despite of being a complex task due to lack of specific vocabulary and difficulty in interpreting the meaning of the posts, sarcasm detection is possible with not 100% accuracy but at least 80% accuracy and can be of important use in the coming future.

References

[1] Bharti, S.K., Vachha, B., Pradhan, R.K., Babu, K.S., & Jena, S.K. "Sarcastic sentiment detection in tweets Streamed in real time: a big data approach", Elsevier 12 July 2016.
[2] Bouazizi, M., Ohtsuki, T., "Pattern-Based Approach for Sar-casm Detection on Twitter" volume 4.
[3] Chaffey, D. Global Social Media Research Summary 2016. URL (http://www.smartinsights.com/Socialmedia-marketing/social-media-strategy/new-globalsocial-media-research/).
[4] Gonzalez-Ibanez, R., Muresan, S., and Wacholder, N. 2011. "Identifying Sarcasm in Twitter: A Closer Look". In Pro-ceedings of the 49th Annual Meeting of Association for Computational Linguistics.
[5] Joshi, A., Sharma, V., & Bhattacharyya, P., "Harnessing Con-text Incongruity for Sarcasm Detection" Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 757-762, Beijing, China, July 26-31, 2015. C 2015 Association for Computational Linguistic.
[6] Maynard, D., Greenwood, M.A. 2014."Who cares about Sar-castic tweets? Investigating the Impact of sarcasm on sen-timent analysis", In Proceedings of the LREC 2014 May 26-31.
[7] Ptacek, T., Habernal, I., & Hong, J. "Sarcasm Detection on Czech and English Twitter", Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 213-223, Dublin, Ireland, August 23-29 2014.
[8] Rajadesingan, A., Zafarni, R., & Liu, H., "Sarcasm detection on Twitter: A behavioral modelling Approach", in Proc. 18th ACM Int. Conf. Web Search Data Mining, Feb. 2015, pp.79_106.
[9] Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N. & Huang, R., "Sarcasm as contrast between a positive sentiment and negative situation", in Proc. Con Empirical Methods Natu-ral Lang. Process, Oct. 2013, pp.704_714.
[10] Tan, W., Blake, M.B., Saleh, I. & Dustdar, S. Social-networkedsourced big data analytics, InternetComput.17 (5) (2013) 62-69.
[11] Veale, T. & Hao, Y. 2010. "Detecting Ironic Intent in Creative Comparisons", In ECAI, Vol. 215.765-770.