Sarcasm Detection on Reddit comments

Lirida Papallazi

University of Milan

January 17, 2021

Outline

Natural Language Processing

NLP enables the computer to interact with humans in a natural manner.

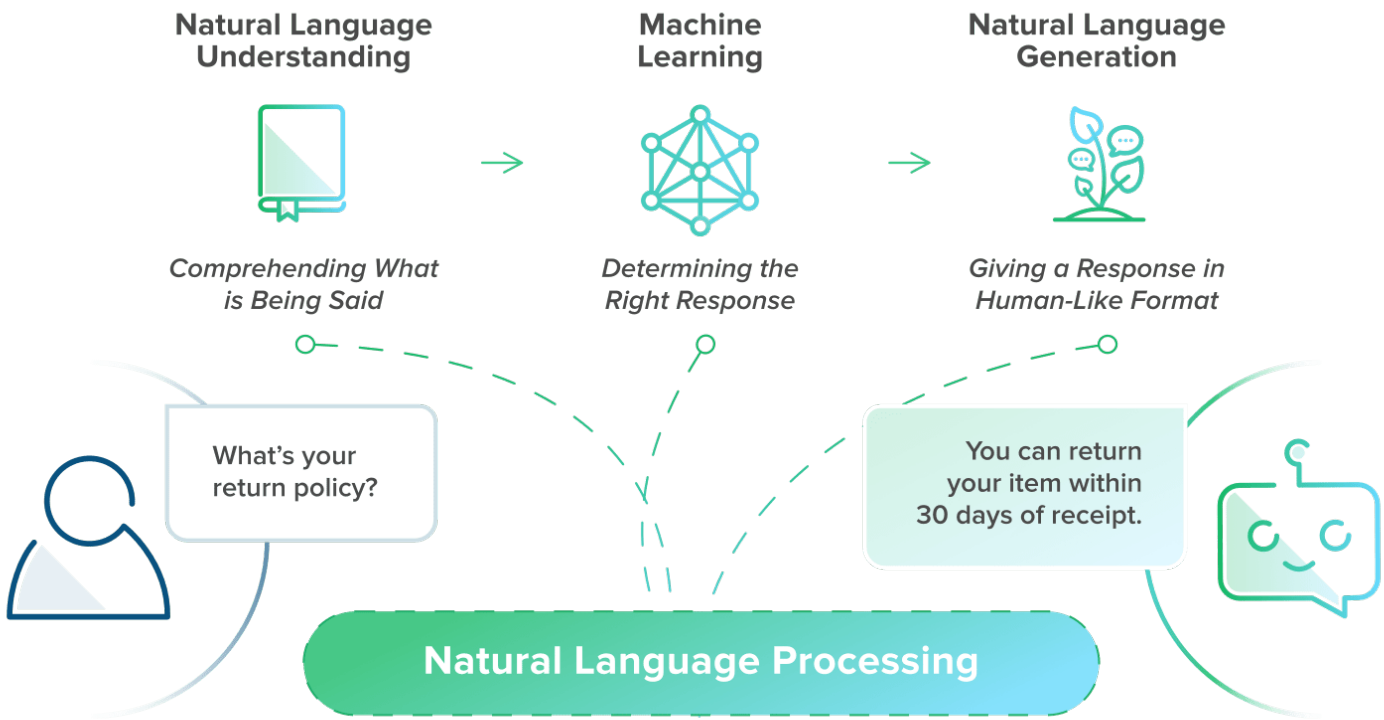


Figure: Schema of the NLP logic

[2]

NLP Techniques

In order to apply the machine learning algorithms the textual data-set has to be pre-pocessed.

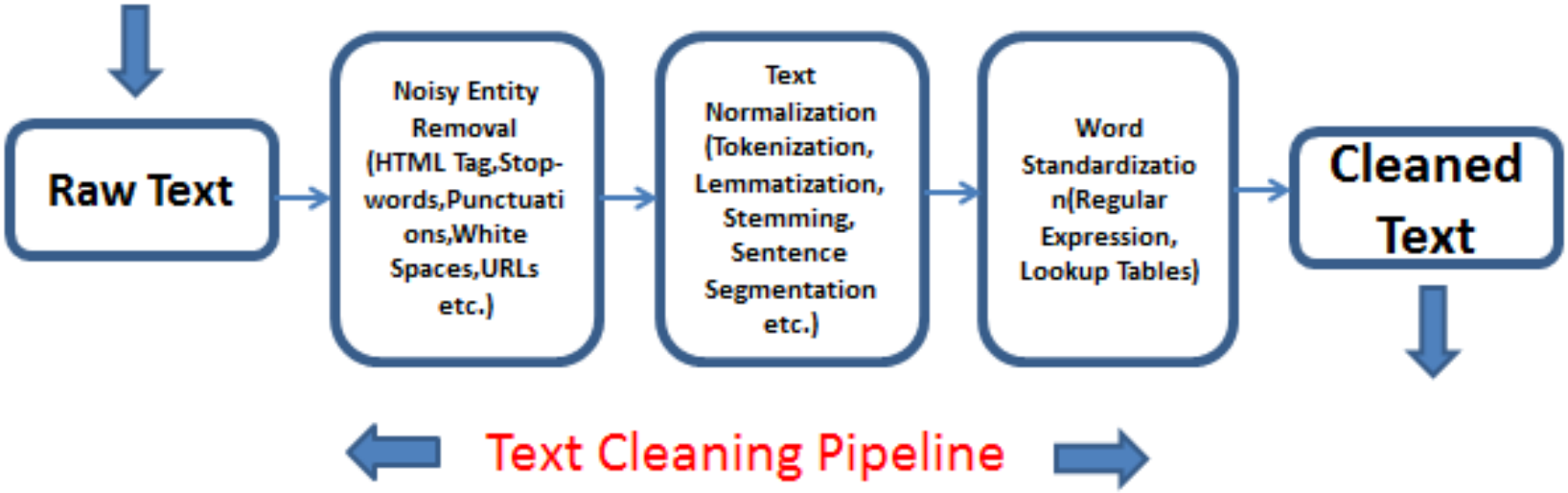


Figure: Text cleaning pipeline

[1]

NLP Techniques (pt. 1)

Stopwords and Punctuation removal Stopwords are typically considered as noise in the text and therefore removed altogether with the punctuation.

Tokenization It is the process of breaking down a text paragraph into smaller chunks such as words or sentence.

NLP Techniques (pt. 2)

Parsing It identifies the structure of the syntax of a text and the dependency relationships between words.

Lemmatization and Stemming NLP uses Lemmatization and Stemming to transform the words back to their root. When the root to which the words are reconnected is the lemma, then we will talk about lemmatization. Stemming instead "trims" the words to obtain the so called stems, which might not be always semantically correct.

Vectorization (pt. 1)

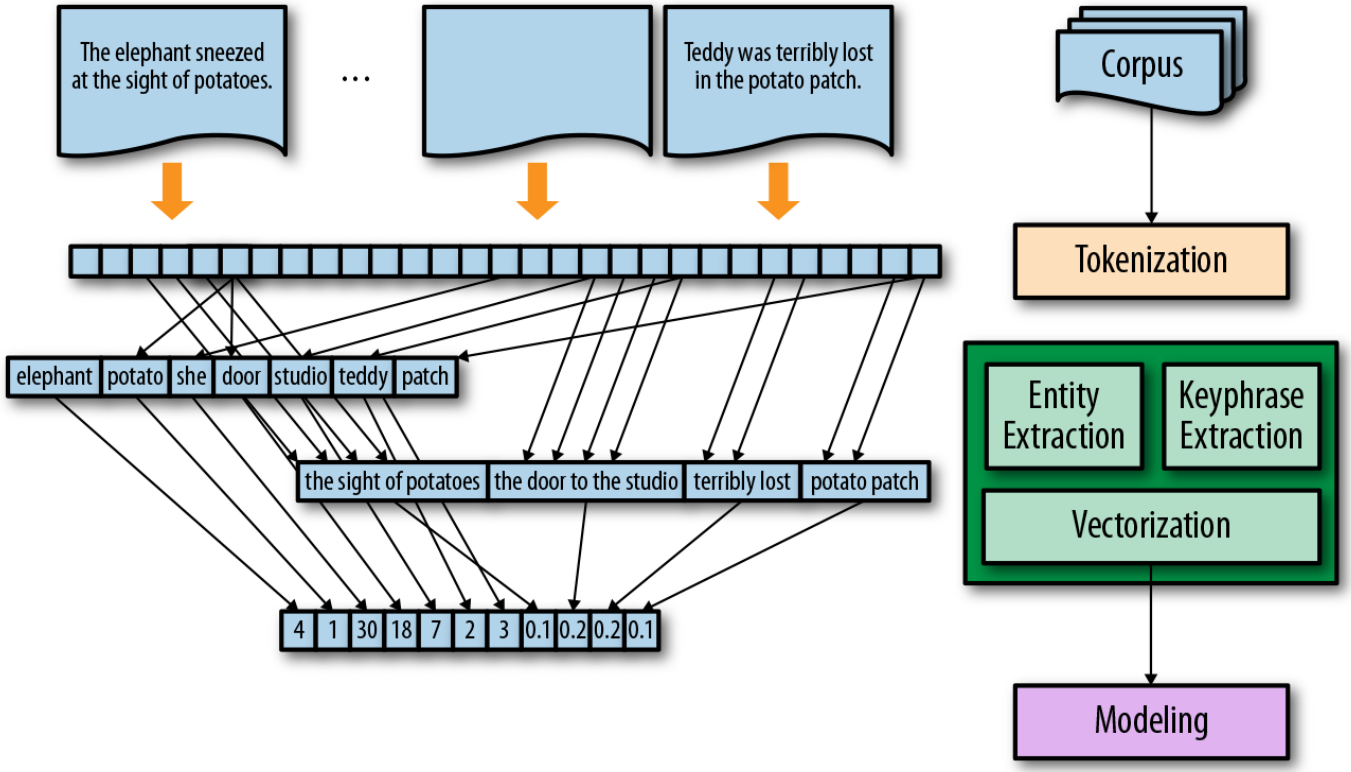


Figure: Process for Bag of Words

[3]

Vectorization (pt. 2)

The application of machine learning algorithms requires that the training data is transformed in vectors, so that the computers can understand the data.

Bag of words Model It is used to keep track of the occurrence of each word;

TF-IDF It is intended to reflect how important a word is to a document in a collection or corpus.

The goal

This project focuses on the analysis of the comments made on the data-set containing the comments made on Reddit, with the aim of predicting the probability that a comment will receive a sarcastic answer, given its subreddit.

What is sarcasm I

According to the Collins Dictionary

Definition

Sarcasm is a noun, speech or writing which actually means the opposite of what it seems to say. Sarcasm is usually intended to mock or insult someone.

but it also is

Definition

- a mocking, contemptuous, or ironic language intended to convey scorn or insult;
- the use or tone of such language.

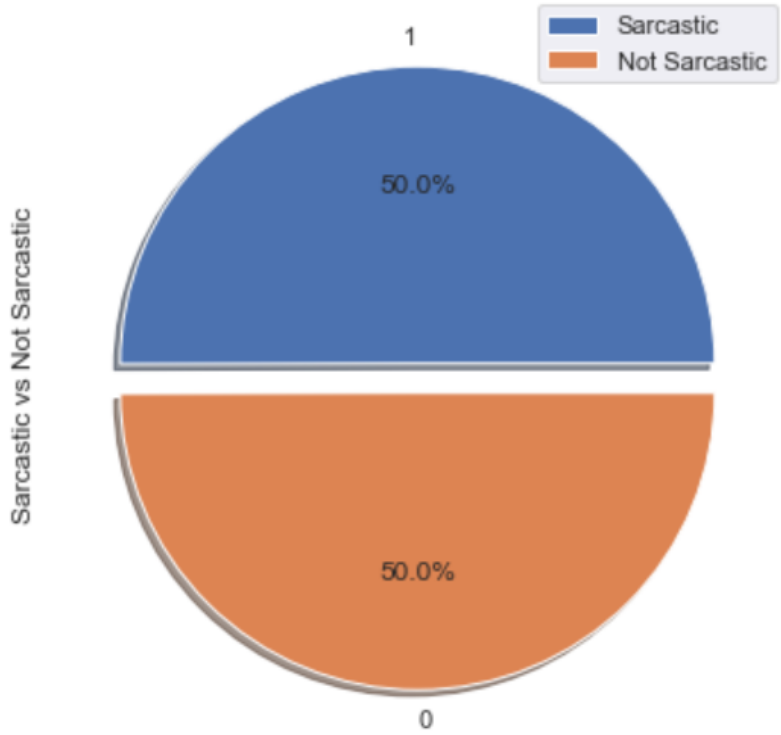
What is sarcasm II

Sarcasm is a communication style that can easily lead to misunderstanding and confusion. Moreover the digital era has transformed the way we communicate with texting, emailing and online commentary replacing face-to-face chats or phone conversations, making it even more difficult to figure out if a writer is being sarcastic.

When delivered in person, sarcasm tends to assume a cutting, bitter tone that in messages don't always pass. Luckily nowadays some solutions such as emojis has been created to mitigate some of the ambiguity.

The dataset

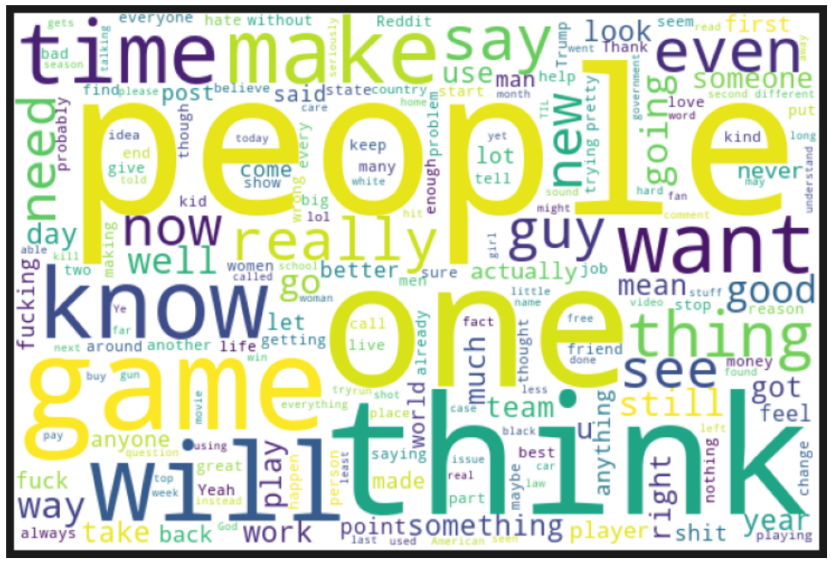
The the data-set is completely balanced in terms of distribution of sarcastic and non sarcastic comments, which means that it does not require any adaptation technique such as oversampling or under-sampling.



Pie chart of the distribution of sarcastic and non sarcastic comments

Wordclouds

A very nice visual representation of text data are Wordclouds, used to depict the most frequently used words in a document.



Wordcloud of parent comments that received a sarcastic answer

Text Classification: logics I

Finally the classification model is built: this model will take the cleaned text of the parent comments as input and it will then come up with a prediction on whether the comment will receive a sarcastic or non sarcastic comment. In particular 3 approaches where used:

Alternative 1 Use two different Tf-Idf vectorizers for the parent comments and that of the subreddits;

Alternative 2 Use both the "cleaned" version of the parent together with the subreddit name as input X, by simply joining the two texts.

Alternative 3 addition to the dataset of the:

- up-votes;
- down-votes;
- count of common words between the parent comment and the comment;

Text Classification: logics II

- number of characters;
- punctuation;
- number of full caps words;
- number of uppercase letters;
- sentiment (categorical variable computed using Vader Sentiment)
- polarity (float value computed using TextBlob)

Text Classification: results I

Different learning algorithms were applied to the two alternatives input data-sets and the following results were obtained.

Classification results for Alternative 1 and 2

Dataset	Naive Bayes	Logistic Regression	Random Forest	Best CV
Alternative 1	57.26%	59.97%	52.94%	NB = 57.46%
Alternative 2	56.82%	60.31%	51.20%	Logit = 57.27%

Text Classification: results II

As for what concerns the third alternatives we report the following results underlining the most meaningful ones.

Classification results for Alternative 3

Dataset	XGBoost	Logistic Regression	Random Forest	KNN
All variables	58.31%	55.69%	52.38%	52.04%
No clean	58.25%	55.79%	56.03%	52.17%
No char	58.23%	57.42%	54.98%	52.94%
No overlap	58.66%	54.87%	51.10%	52.25%
No cap	58.96%	55.17%	54.36%	52.34%
No punct	58.68%	55.20%	53.50%	52.38%
No upper	58.67%	55.83%	53.71%	52.57%
No ups	56.96%	55.42%	53.37%	52.17%
No downs	58.16%	55.31%	51.18%	52.13%
No polarity	58.65%	55.87%	54.38%	52.45%

Let's try with an example

We tried to compute the probability of receiving a SARCASTIC answer of a comment taken directly from Reddit and with AskReddit as subreddit, using only the first and the second alternative.
The comment is "What is your dream job? (Serious)".

Dataset	Naive Bayes	Logistic Regression
Alternative 1	47.095%	44.36%
Alternative 2	25.71%	18.87%

Conclusion I

As it is possible to observe the results cannot be considered the best ones: the best accuracy achieved barely reaches the 60%. This might be a consequence of four major facts:

Indirect approach We are using an indirect approach to discover the probability of receiving a sarcastic answer and the labels actually refer to the comments. We proved that running the same algorithms on the comments leads to better performances, confirming this possibility.

Questionable labelling Analyzing some of the comments and their labels it is possible to note that some obviously sarcastic answers have not been labeled as such.

Size matters All the analysis was performed on a very small sample of the original data-set. However even changing the sample size, the results don't seem to significantly change.

Conclusion II

What is sarcasm? The issue with the aim of this project is the aim itself. Even though there may be some features that increase the chances of receiving a sarcastic answer it is possible to answer to any kind of message or sentence with a sarcastic comment, hence our results and in particular the classifiers' accuracy when the parent comments text is taken out of the equation, seems compatible with this observation.

Conclusion: Indirect procedure

In order to prove this cause we run the same exact code and procedure but with focus on the children comments. It is possible to observe from the table, there are some significant improvements meaning that this is a more the plausible cause.

Classification results for children comments

Dataset	Algo 1	Logit	Random Forest
Alt 1	NB: 64.26%	58.15%	70.85%
Alt 2	NB: 65.08%	71.32%	56.55%
Alt 3	XGBoost: 62.46%	57.84%	55.83%

Conclusion: What is Sarcasm I

We plot here the Top 10 most "important" words or features that impact on sarcasm for respectively the analysis on the parent and on the children comments.

Conclusion: What is Sarcasm II

0	assault
1	bear
2	billion
3	birthday
4	black
5	bomb
6	buck
7	butter
8	cat
9	children

Figure: Parent comments

0	add
1	amirit
2	black
3	clear
4	dare
5	drop
6	duh
7	everyon
8	everyon know
9	fault

Figure: Children comments

The End

References


Vanaudel Analytix.
Pre-processing text for nlp (24/08/2019).


Bold360.
A complete guide to natural language processing (nlp).


O'Reilly.
Chapter 4. text vectorization and transformation pipelines.