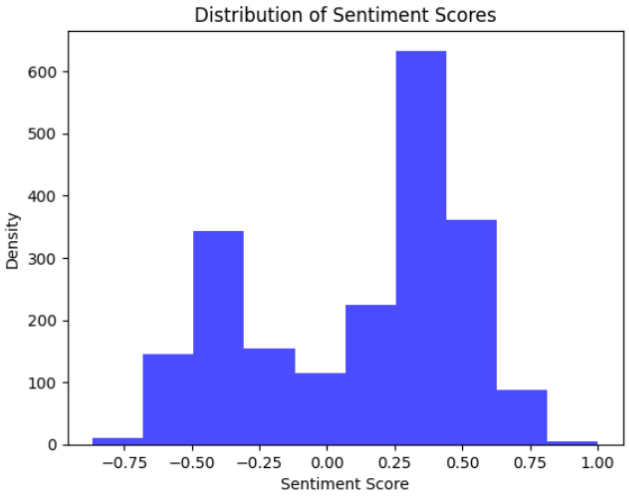
Exploratory data analysis

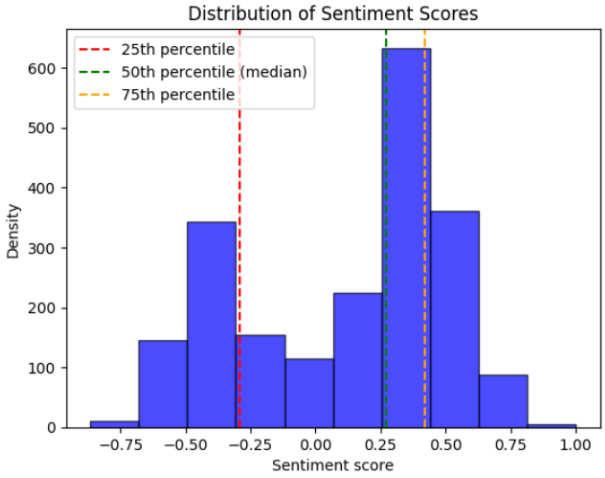
Semval2017

The Semval2017 dataset consists of 2078 instances of microblogs, structured in a JSON file format. Each microblog entry includes an attribute termed "sentiment score," denoting the sentiment associated with the respective microblog post. The distribution of sentiment scores is visually depicted in the accompanying plot. Within the dataset, 710 entries exhibit sentiment scores below 0, while 31 entries possess a sentiment score of 0. Furthermore, the dataset encompasses 1337 entries with sentiment scores exceeding 0.

Upon analysis, it was determined that the microblog entry with the longest character span contains 100 characters, while the entry with the highest word count comprises 20 words.



The next plot describes also the distribution of data in quartiles.



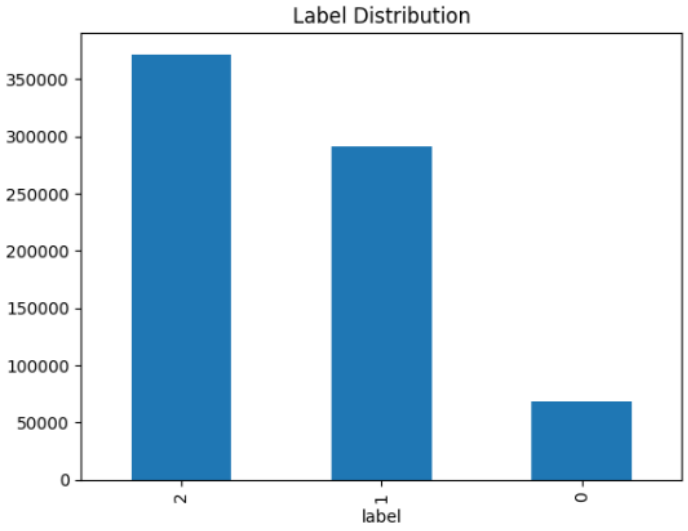
Examples:

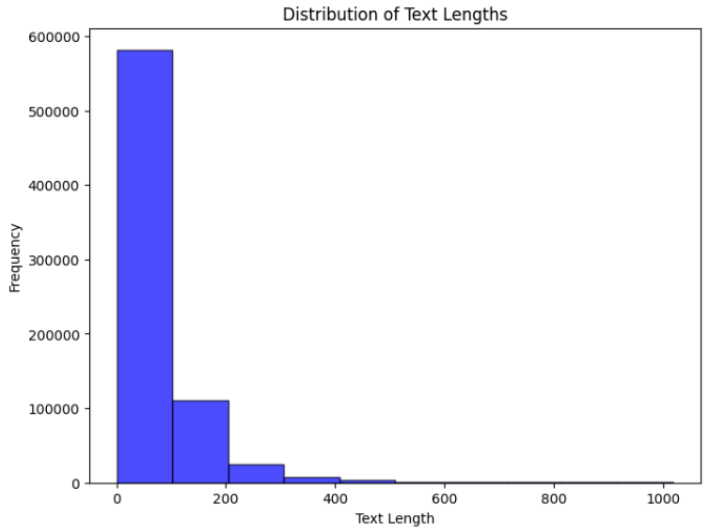
{
"source": "stocktwits", "cashtag": "$SPY",
"sentiment score": "**-0.850**",
"id": "6181018",
"spans": ["**look out belooooooowwww**"]
}

{
"source": "stocktwits", "cashtag": "$SPY",
"sentiment score": "**0.829**",
"id": "6058084",
"spans": [**"Yes buy everything up**"]
}
{
"source": "stocktwits", "cashtag": "$VRX",
"sentiment score": "**0.000**",
"id": "11570386",
"spans": ["**Another quiet week**"]
}

Stocktwits-crypto dataset

The Stocktwits-crypto dataset contains 731 697 instances of twits stored in CSV format, comprising of textual data paired with corresponding sentiment labels. The initial plot illustrates the distribution of text labels. The subsequent plot describes the distribution of text lengths. The maximum text length observed is 1017 characters, while the maximal word count within a single twit amount to 301.





Examples:

Label 2 : it’s our time to glitter in the mines! what time is the show?

Label 0 : musk selling his bitcoins, he need money, he will be broke soon

Label 1: are steals and going lower if you got the spare cash excellent long term buy and holds rn for bulls and only getting better