Done

In combine data.csv

1st: go through data i.e column C (text column) and check if apple or amazon or both or neither

If neither = 0

If amazon = 1

If apple = 2

If both = 3

Result goes in E

Done

2nd: do semantic analysis on Column C (text column) and check if positive or negative article

Range for result: (-1, 1)

If negative: negative integer

If positive: positive integer

Result goes in F

Done:

3rd: iterate through column B and keep track of each news and number of times it posted

Ie. Column B: total times it posted article

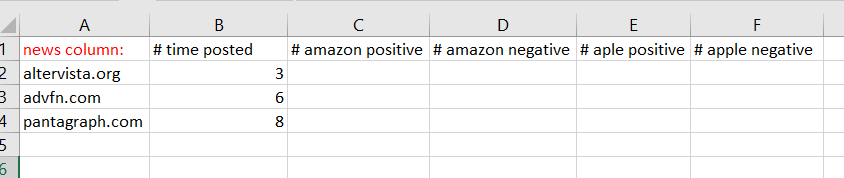
Column C: positive amazon

Column D: negative amazon

Column E: positive apple

Column F: negative apple

Column G: rank news source



4th: Accuracy News

How accurate was their news to the true stock price change?

News correct: semantic news (column F) reflects stock price

Total1: # = (news correct/total news from that site)

Avg news posted (anp): (Total Column B/n)

Weight the news by what time it was published i.e if it was published by peak time or dead time

Hsst: Hour average of stock trade

avg val (g): Column G (total stock)

Total2: #= (hsst/g)

Total3: (Total1+Total2)/anp

Volatility of hour: take each hour and see the change in delta

(Column C sum – Column F sum)/n

To get the avg delta per hr

So more volatile hour should be weighted more than others

??????????????

Give rank for news in crucial time i.e more points if posting in peak time