Investigating Solar Cycles and Media Sentiment Through Advanced NLP Techniques

Aghamusa Azizov

Abstract**—**This study investigates the correlation between solar activity and sentiment in news media coverage, using a large-scale dataset of solar activity since 1750 and over 15 million articles from "The New York Times" dating from 1851 onwards. Employing Pearson's correlation coefficient and multiple Natural Language Processing (NLP) tools—TextBlob, Vader, and DistillBERT—the research examines the extent to which fluctuations in solar phenomena are reflected in the sentiment of historical news narratives. The findings reveal that the correlation between solar activity and media sentiment is generally negligible, suggesting a weak influence of solar patterns on the portrayal of events in news media. Notably, a moderate positive correlation was observed between the sentiments derived from TextBlob and Vader, indicating consistency across NLP tools. The analysis provides insights into the historical impact of solar activity on human affairs and highlights the importance of using multiple analytical methods to understand complex relationships in large datasets. The study contributes to the broader understanding of how extraterrestrial factors may intersect with media-reported events and underlines the intricate nature of interdisciplinary research in the data science and historical domains.

Keywords**—**Solar Activity Correlation, Media Sentiment Analysis, Natural Language Processing, NLP, Historical Event Patterns.

I. Introduction

HUMANITY has always sought periodic patterns in nature and the surrounding environment, linking them to specific life events in an attempt to predict occurrences, excesses, and calamities. This work attempts to find a connection between solar activity and certain events on Earth. Initially, German astronomer and botanist Heinrich Schwabe observed sunspots and made drawings of them from 1825 to 1867, suggesting in 1838 a potential ten-year cycle of sunspot activity [1]. Following this, Swiss scientist Rudolf Wolf was greatly impressed by Schwabe's discovery of the sunspot cycle and not only conducted his own observations but also compiled all available data on sunspot activity dating back to 1610, calculating an 11.1-year period for the cycle [2]. In 1848, he devised a method to quantify sunspot activity, known today as the Wolf number, which remains in use. In 1852, Wolf was among four individuals who discovered the link between the cycle and geomagnetic activity on Earth [3].

Subsequent efforts by scientists across various fields have sought connections between solar activity and their areas of research. One of the first to notice this connection was Nobel Prize laureate in Chemistry, Swedish physical chemist Svante Arrhenius, followed by Russian-Soviet scientist Alexander Chizhevsky, a pioneer in several scientific disciplines related to space, including cosmic natural science, cosmic biology, ecology, and epidemiology [4], [5]. Chizhevsky identified links between solar activity and numerous Earth events, such as wars, revolutions, and epidemics. However, Chizhevsky later faced dismissal in the USSR for falsifying results, and his scientific activities were condemned as charlatanism, despite his continued invitations to preside over international space-related conferences [6].

Numerous scientists in various fields have observed the connection between the periodicity of solar activity and various phenomena and processes on Earth. American scientist Douglas noted that solar activity affects the growth rate of trees [7], [8], while Soviet entomologist Shcherbinovsky found that the periodicity of locust plagues corresponds to the 11-year solar cycle [9]. Hematologist Shultz discovered that fluctuations in solar activity affect the number of leukocytes in human blood and relative lymphocytosis [10]. Italian physical chemist Piccardi observed the influence of physical factors, including solar activity changes, on colloidal solutions [11]. Japanese hematologist Takata developed a blood protein sedimentation test sensitive to solar activity changes [12]. Additionally, a series of American economists noted a correlation between solar activity and agricultural production, as well as business activity curves [13], [14].

Chizhevsky conducted a historiometric analysis of sunspot records, comparing them to significant historical events in Russia and 71 other countries from 500 BCE to 1922 CE. His findings revealed that a considerable percentage of major historical events involving large groups of people occurred around sunspot maximum. Edward R. Dewey, founder of the Foundation for the Study of Cycles, analyzed and published Chizhevsky's data in the Foundation's publications in 1951 [15]. In [16], Dewey explained the "four components" of Chizhevsky's eleven-year cycle and their estimated durations: a three-year period of minimum activity marked by passivity and autocratic rule, a two-year period where the masses begin to organize under new leaders and a single theme, a three-year period of maximum excitability, revolution, and war, and a three-year period of gradual decrease in excitability until the masses become apathetic. While Dewey questioned Chizhevsky's theory due to the lag time between sunspot cycle height and his "mass excitability index," his contributions to the field remain significant.

In [17], Putilov empirically tested the Chizhevsky hypothesis by analyzing events described in Soviet historical handbooks. He discovered that the frequency and "polarity" of

A. Azizov is prospective Computer/Data Science student with Baku, Azerbaijan (e-mail: agamusa.aziz@gmail.com).

Open Science Index, Aerospace and Mechanical Engineering Vol:18, No:9, 2024 publications.waset.org/10013800.pdf

events, including revolution, is highest in the years of the solar cycle maximum and lowest in the year before the minimum.

The purpose of this study is to explore the possible link between patterns of solar activity and the mood or tone of news reporting over an extensive period. By examining this potential connection, the authors seek to address a long-standing question: Can changes in solar activity influence how events are reported and, by extension, perceived by the public? The study leverages advanced data analysis tools to assess sentiment in historical news articles against solar activity data. This approach not only endeavors to clarify the relationship between space weather and human sentiment but also to refine the analytical techniques used in such investigations. The findings are intended to inform a more grounded perspective on how natural phenomena may subtly impact societal narratives and to set a precedent for future research at the intersection of environmental science and social studies.

II. Methodology

The first dataset utilized for the correlation analysis was a Solar Activity dataset starting from the year 1750, acquired from Kaggle. This dataset provided comprehensive details on sunspot observations, which are crucial for understanding the long-term patterns of solar activity. Media information used for correlating with Solar Activity was derived from the archives of "The New York Times," encompassing over 15 million articles, including abstracts and headlines, collected from the year 1851 to the present. This extensive collection served as a valuable resource for analyzing the influence of solar activity on public sentiment as reflected in news coverage.

To conduct the correlation analysis between solar activity and sentiment analysis results from news media, the study employed three distinct sentiment analysis methods:

1. TextBlob: Identified as a straightforward, traditional, and rapid NLP tool, TextBlob offers functionalities for sentiment analysis, spelling correction, and more. For this study, sentiment analysis was conducted using the NaiveBayesAnalyzer polarity parameter, which provided sentiment values. The sentiment values derived from the analysis of articles ranged from 0.009 to 0.199.
2. VADER (Valence Aware Dictionary for sEntiment Reasoning): VADER is a rule-based sentiment analyzer known for its effectiveness in identifying the sentiment orientation of lexical features within a text, such as words labeled as positive or negative. The sentiment scores for the analyzed articles varied from -0.0899 to 0.18.
3. DistilBERT (Distilled Bidirectional Encoder Representations from Transformers): As the most advanced NLP model for sentiment analysis mentioned in this study, DistilBERT represents a streamlined version of the original BERT model, optimized for faster performance while retaining a significant portion of its predecessor's effectiveness. The sentiment analysis scores for articles utilizing this model ranged from -0.3251 to 0.2865.

Despite initial plans to solely employ DistilBERT analysis via the HuggingFace platform, practical constraints necessitated exploring alternative methodologies for local sentiment analysis. This approach involved the use of a downloaded model file (sentiment-en-mix-distillbert_4.pt), significantly reducing analysis time per text and enabling parallel processing across multiple computers without the periodic limitations encountered on the HuggingFace platform.

The study's findings, illustrated in Fig. 1 of solar activity, underscore the presence of 11-year cycles in solar activity, aligning with historical records and providing a basis for further examination of their correlation with global events and media sentiment.

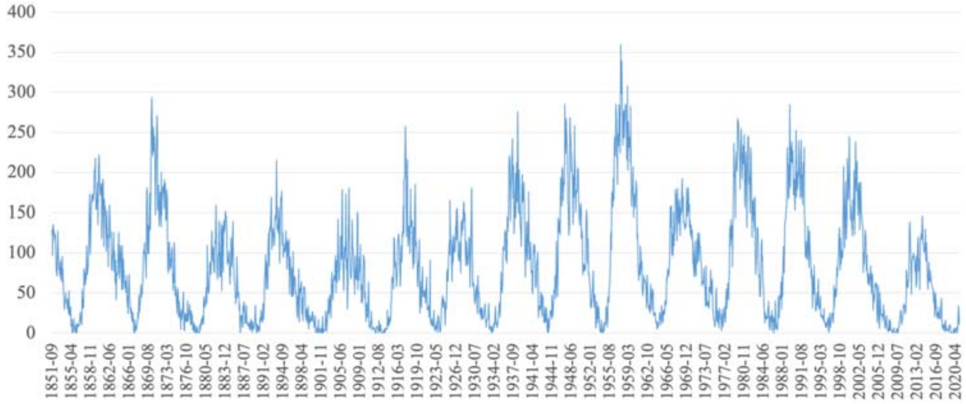

Fig. 1 Solar intensity variations between 1851-2019 by the Kaggle dataset

III. Discussion

The Pearson correlation coefficient formula, calculated in Excel, is as follows:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where r is the correlation coefficient ranging from -1 to +1, n is the number of samples (texts for analysis), about 15 million, x, y are pairs of input data used for the correlation analysis, for example, Solar-TextBlob, Solar-Vader, Solar-Flair(DistillBert), as well as TextBlob-Vader, TextBlob-Flair, and Vader-Flair.

The results of the analyses are presented in Table I.

TABLE I
SUMMARY OF THE CORRELATION RESULTS FOR DIFFERENT NLP ANALYSES

	Solar	TextBlob	Vader	Flair (DistilBert)
Solar	1			
TextBlob	0.021	1		
Vader	0.183	0.621	1	
Flair (DistilBert)	0.111	-0.445	-0.219	1

The analysis revealed a very low correlation coefficient between solar activity and events on Earth. Based on [18], the correlation coefficients "r" can be considered negligible correlations between solar activity and each of the sentiment analysis methods. The correlation coefficient "r" is very small between solar activity and all three sentiment analysis methods:

- Solar-TextBlob = 0.02,
- Solar-Vader = 0.18,
- Solar-Flair (DistilBert) = 0.11,

Overall, the results are indicating a weak link between solar activity and news, and accordingly, with events on Earth. However, the fact that sentiment analysis accurately determines the tone of the news is shown by the mutual correlation of two methods, TextBlob-Vader, for the same texts r = 0.62 (highlighted in yellow in the table), corresponding to a moderate positive correlation (in the range of 0.5-0.7) [18]. In contrast to the negligible correlation and the TextBlob-Flair result of -0.44 (Low Negative Correlation), it indicates some small correlation.

Fig. 2 shows the dynamics of solar activity and the three NLP sentiment analyses, preliminarily multiplying the NLP analysis values by 1000 (TextBlob1K, Vader1K, Flair1k) to visually bring them to the same dimension. However, the results in the figure does not indicate a dependency between solar activity and news in the media.

Perhaps all the dependencies listed above found by past scientists are related to the fact that they took correlations not for the entire period from 1851 to 2023, but for certain limited periods of time, during which even visually (according to Line Charts) a connection and synchronicity of peaks and troughs can be seen.

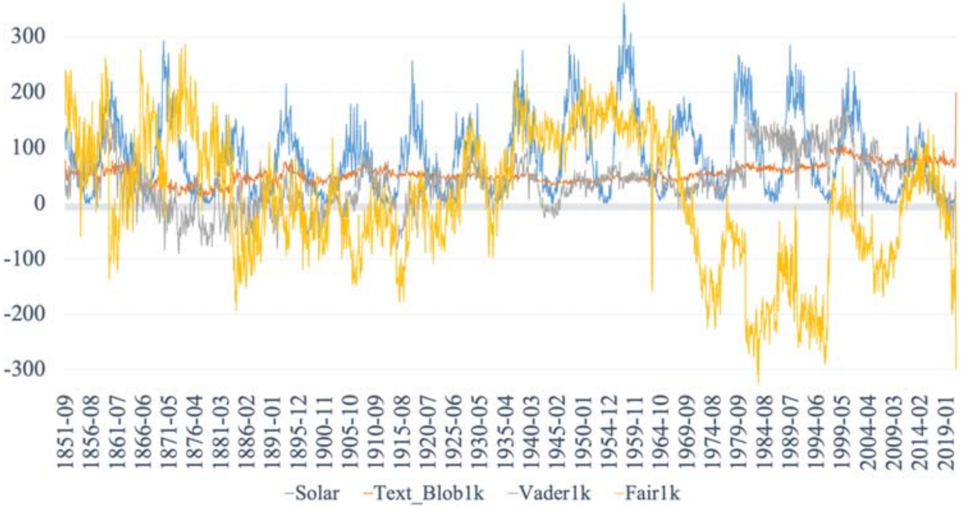

Fig. 2 Diagram of correlation between solar activity and sentiment analysis coefficients calculated by different NLP processing methods for New York Times news articles

In an effort to explore potential connections between solar activity and events on Earth, this study also conducted a factor analysis between the dependent variables (outcomes of sentiment analysis through various methods) and the independent variable (solar activity). Scatter plots for each pair, characterized by elongated clouds (Figs. 3-5), indicate the homogeneity of observations or conditions of homoscedasticity (consistent variance within the model of linear regression), thereby enabling us to perform a factor-regression analysis on the data pairs (Solar-TextBlob, Solar-Vader, Solar-Flair).

The coefficient of determination of linear regression results presented in Figs. 3-5 is computed in Excel using the following expression:

$$R^2 = 1 - \frac{SSreg}{SStot}$$

where R^2 is Coefficient of determination; SSreg is Sum of squares of residuals; SStot is Total Sum of squares.

IV. CONCLUSION

The comprehensive correlation analysis conducted between solar activity and media sentiment, represented through various NLP methods, yields a nuanced understanding of the potential relationships between astronomical phenomena and terrestrial events as reported in news media. The investigation, grounded in Pearson's correlation formula and robust datasets spanning from 1851 to the present, reveals predominantly negligible correlations between solar activity and sentiment scores obtained from more than 15 million "The New York Times" articles.

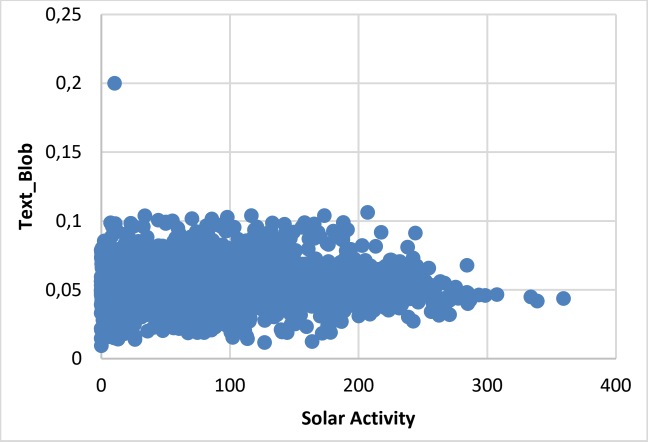

Fig. 3 Scatter plot for Relations between Sentiment analysis (TextBlob) and Solar Activity

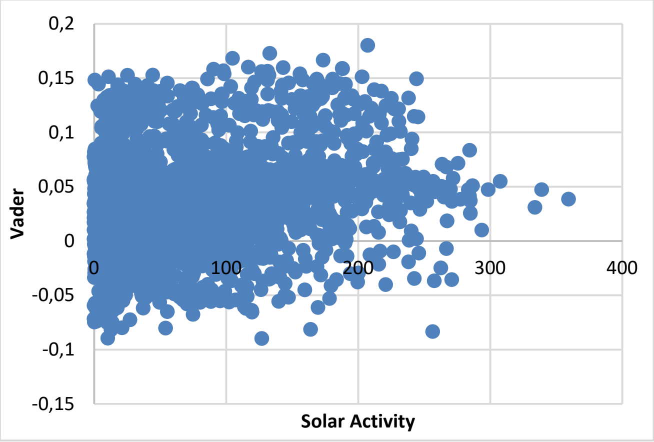

Fig. 4 Scatter plot for Relations between Sentiment analysis (Vader) and Solar Activity

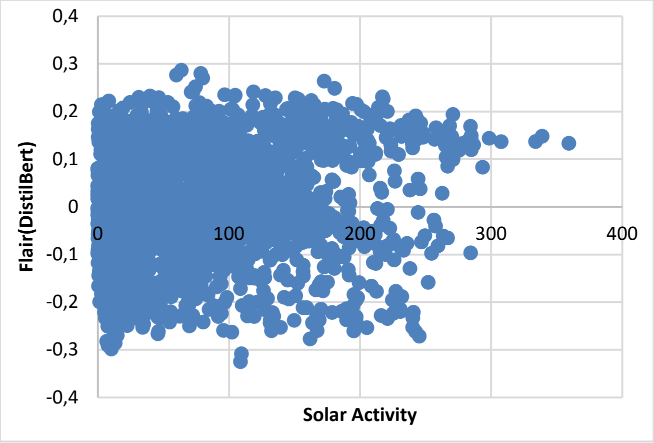

Fig. 5 Scatter plot for Relations between Sentiment analysi s(Flair) and Solar Activity

TABLE II
RESULTS OF REGRESSION ANALYSIS FOR SOLAR-TEXTBLOB PAIR

Multiple R	0.020190993
R Square	0.000407676
Adjusted R Square	-8.49765E-05
Standard Error	68.1779109
Observations	2031

TABLE III
RESULTS OF REGRESSION ANALYSIS FOR SOLAR-VADER PAIR

Multiple R	0.183080606
R Square	0.033518508
Adjusted R Square	0.033042174
Standard Error	67.03922795
Observations	2031

TABLE IV
RESULTS OF REGRESSION ANALYSIS FOR SOLAR-FLAIR PAIR.

Multiple R	0.11072422
R Square	0.012259853
Adjusted R Square	0.011773042
Standard Error	67.77251251
Observations	2031

The correlation coefficients between solar activity and the sentiment analysis methods—TextBlob, Vader, and Flair(DistilBert)—are notably low (0.02, 0.18, and 0.11, respectively). These results suggest a weak connection between solar activity and media sentiment, challenging the notion that solar phenomena have a direct and significant impact on the discourse within economic, political, or medical contexts as conveyed through news reporting. However, the moderate positive [18] correlation (0.62) found between TextBlob and Vader sentiment analysis methods indicates a reasonable consistency in the sentiment conveyed by these two distinct NLP tools when applied to the same textual data.

The divergence in correlation between TextBlob and Flair(DistilBert) underscores the variability in sentiment analysis outcomes, depending on the chosen NLP method, and suggests the presence of a low negative correlation. This disparity emphasizes the need for a diversified approach when employing NLP tools for sentiment analysis to ensure a comprehensive understanding of textual data.

The determination coefficient R^2 (R Square in Excel), derived from regression analysis for the three pairs of values Solar-TextBlob, Solar-Vader, and Solar-Flair, indicated a very weak correlation between solar activity and sentiment analyses conducted by three methods—0.000407676, 0.033518508, and 0.012259853, respectively. These values, being less than 0.5, suggest the insufficiency of the mathematical regression model to adequately describe the relationship between terrestrial events and solar activity.

The visual analysis, though employing line charts to illustrate the parallel trends of solar activity and sentiment analysis over an extended period, does not visually corroborate a significant relationship between solar patterns and news content. This observation aligns with the statistical findings and leads to the conclusion that if any relationship exists, it may manifest over specific, limited time frames rather than across extensive historical periods.

In light of these findings, the study posits that the correlations identified by past researchers may have been influenced by the specific time intervals selected for their analyses, which may have been more visually apparent in their synchronicity of peaks and troughs. This highlights the complexity of establishing a definitive link between solar activity and human affairs and points to the need for further nuanced and targeted research in this domain. The exploration of such correlations must consider the multifaceted nature of human activity and the multitude of variables at play, beyond the scope of solar influences.

REFERENCES

[1] Schwabe, H. (1844). Sonnenbeobachtungen im jahre 1843. von herrn hofrath schwabe in dessau. Astronomische Nachrichten, volume 21, issue 15, p. 233, 21, 233.

[2] Wolf, R. (1852). Neue untersuchungen über die periode der sonnenflecken und ihre bedeutung.. *(New investigations regarding the period of sunspots and its significance). Mittheilungen der Naturforschenden Gesellschaft in Bern (in German). 255: 249–270.*

[3] Wolf, R. (1852). Sonnenflecken-Beobachtungen in der ersten Hälfte des Jahres 1852; Entdeckung des Zusammenhanges zwischen den Declinations-variationen der Magnetnadel und des Sonnenflecken. Mittheilungen der Naturforschenden Gesellschaft in Bern, 179-184. Mittheilungen der Naturforschenden Gesellschaft in Bern (in German). 245: 179–184.

[4] Tchijevsky, A. L. (1971). Physical factors of the historical process. Cycles, 22, 11-27.

[5] Chizhevsky, A. L. (1930). Epidemic catastrophes and periodic activity of the Sun. Moscow (in Russian).

[6] “On the dismissal of the director of the central laboratory for ionification A. L. Chizhevsky” (July 8, 1936) Pravda newspaper.

[7] Douglass, A. E. (1933). Tree growth and climatic cycles. The Scientific Monthly, 37(6), 481-495.

[8] Douglass, A. E. (1927). Solar records in tree growth. Science, 65(1679), 220-221.

[9] Shcherbinovsky N. S. (1960). Solar-induced cyclicity of mass reproduction of harmful insects and other animals. - Astronomer. Sat., issue 3/4, p. 165-169

[10] Shultz N. A. (1964) The influence of solar activity fluctuations on the number of white blood cells. - In the book: Earth in the Universe. Moscow (in Russian): Mysl, p. 382-399

[11] Pikkardi G. (1962) The chemical basis of medical climatology. - USA: Springfield, - 146 p.

[12] Takata, M. (1951). Über eine neue biologisch wirksame Komponente der Sonnenstrahlung: Beitrag zu einer experimentellen Grundlage der Heliobiologie. Archiv für Meteorologie, Geophysik und Bioklimatologie, Serie B, 2, 486-508.

[13] “The Quarterly Journal of Economics” (1934) November

[14] Garcia-Mata, C., & Shaffner, F. I. (1934). Solar and economic relationships: a preliminary report. The Quarterly Journal of Economics, 49(1), 1-51.

[15] Journal of cycle research, Volumes 1–5, p 89 (1951) and Volume 41, p 156

[16] Dewey, E. R., & Mandino, O. (1971). Cycles: The mysterious forces that trigger events. (No Title).

[17] Putilov, A. A. (1992). Unevenness of distribution of historical events throughout an 11-year solar cycle. (Article in Russian) Biofizika, 37(4), 629-635.

[18] Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal, 24(3), 69-71.