**REVISITING THE CONNECTION BETWEEN SOLAR ERUPTIONS AND PRIMARY HEADACHES AND MIGRAINES USING TWITTER**

This paper aims to find a connection between the solar eruptions and the primary headaches and migraines using twitter data. The results proved that the two are not correlated to each other. For the study the author collected over 7.1 million tweets where majority of the tweets reported were actual, real-time symptoms. The strong point about the results of the study is that in order to determine the confidence interval of their null result the sensitivity of the data and the method the author injected an artificial signal in the data and the results remain unchanged showing the correctness of the study. The author also states that children or seniors who are not well represented in the data may be more susceptible to solar activity as a trigger. There are a lot of weak points in this study: (1) the author had no information about the severity of the symptoms. (2) there is a lack of studying the offline behavior of the study carried out as a question arises “Why would anyone want to tweet during a headache or Migraine?”. (3) the author did not filter out the tweets using geo-tags as people closer to the poles might be more affected to such a phenomenon rather than considering the tweets of the people at the poles. (4) The only keywords used by the author were “head-ache[s]”, “head-ache[s]”, “migraine[s]” which is not enough for such a complex study as people might use other synonyms or terms even in their regional languages. (5) The author removed all the re-tweeted message rather than considering them, as if a message is retweeted more people may be experiencing such problem.

**UNDERSTANDING U.S. REGIONAL VARIATION WITH TWITTER DATA ANALYSIS**

The paper aims to understand the difference in the linguistics among various counties across the U.S using twitter data. To carry out this study the author used a huge dataset of 7.8 billion words and 6.6 million twitter users. The author also details out a rich background history of this study and techniques which were previously used. Some of the previous techniques outlined were Natural Language Processing, Parts of Speech Tagging and Latent vector autoregressive model this was used to study the diffusion of linguistic change over the united states. Multivariate analysis and principle component analysis were some of the statistical methods that were used before. The author used a method of lexical alternation to carry out his study which is basically a bag of synonyms for a particular word and then selecting the top 2 from them. Out of the 211 alternation the author eliminated 152 infrequently used alternations on the basis of if its not present in more than 1000 counties. The author then calculated variant preference (VP) and mean-variant-preference(MVP). The alternations were then selected using spatial autocorrelation testing. Now since the MVP’s were not reliable the author used adaptive kernel smoothing for each alternation. The next step was using principle component analysis of 38 linguistic alternations to extract orthogonal dimensions and then use multivariate regionalization method to detect natural hierarchy of dialect regions in U.S. With 13 input variables the author constructed a 2d multivariate map with similar clusters that are nearby and colored them on a particular scale. These produced highly meaningful results. However questions asked would be (1) tweets about migrants? Can they be filtered out from the native county residents are their linguistics would differ (2) The authors assumption that misspelling and abbreviations are not a major concern to the final result.