Sentiment Analysis Using Twitter Data

Kenneth Xavier Dsilva
ID: 10600644
10600644@mydbs.ie

Handup Date: 23/06/2022

1 Content Introduction.

Sentiment Analysis on a Twitter data-set in the present day has been found to be extremely repetitive and repulsive. However, many breakthroughs in the field of science have been through repetition of everyday activities with just one variable- "Curiosity". This, I say, is based on the infamous story of The apple and Einstein. Hence, my curiosity led me to critique one of the starting points of this explosive development in Natural Language Processing and its usage on Twitter in particular. The paper, I critique, was published to the IEEE community in 2019 by Md. Rakibul Hasan, Maisha Maliha, and M. Arifuzzaman under the title 'Sentiment Analysis with NLP on Twitter Data' (Hasan, Maliha, and Arifuzzaman 2019) My focus throughout this critique won't be concentrated on judging the efforts through the spec- trum of technological boundaries as in 2019 or an inert comfort towards conventional approaches, but also towards the end goal and their vision with the results. The team has generated a great accuracy of 84.25% through their Logistic Regression algorithm with fine tuning in the data prepossessing state. However, the use of their model on real-time data or their confinement in the area of tokenization will be discussed in detail ahead.

2 Influence of the Paper and Finding the Gap.

2.1 Objective.

The impact of this paper was majorly on the comparison of individual influence over a set of prejudices and choices. The authors could firmly distinguish and state that a general influence on social media could be traced back to the sales and success of a certain brand of mobile providers. The team compared two leading mobile service providers, namely Apple and Samsung. The popularity was based upon tweets with negative feedback about the particular brand. A total of 1000 tweets for both iPhone and Samsung phones are analyzed. However, they are also divided into batches of 50, 100, 500, and a maximum of 1000 reviews for each. Graphs regarding the feedback and categorization are also provided to make visualization easier.

2.2 Questions Answered and Methodology Used.

The team have cross referenced their work over different algorithms, namely Naive Bayes Algorithm (accuracy: 82.1%), K-Nearest Neighbors (accuracy: 74.74%) and finally proved their success with Logistic Regression with Support Vector Machine, giving an accuracy of 85.25%. The major contribution in the methodology, however, wasn't the application of the model but the pre-processing of the data and the steps taken to do so. Their exploitation of a significant term resulted in the creation of an excellent model, and that was the min df and max df. These factors, in simple words, control the tokenization and repetition of the words tokenized. To put it in simpler words, the higher the min DF, the more reduction of less frequent tokenized words. Hence, the entire prospect of the 85.25% accuracy depends upon the frequency of the tokenized words, which honestly does not support a constant accuracy in real-time data.

2.3 Findings and Results analysis.

The team's process for reaching a conclusion and delivering their outputs is very simple and straightforward. However, there are many loopholes that can affect the sanity check of the result interpretation. The positive to negative graphs are helpful to generalize the sentiment of the tweets and conveniently say that the iPhone(Apple) has higher positive and lower negative feedback compared to Samsung. However, no firm comparison to any such statistics is provided to actually prove that IPhone is better chosen than Samsung. The results are said to hold true to the fact that the iPhone is a better seller or a better Mobile device provider than Samsung. This is purely done using the results provided by the algorithm and the 1000 tweets. With an accuracy of 85.25%, one might believe that this result is very true and, on the first look, it does seem that way. However, with a deeper understanding of the approach used, it could be very different.

3 Critique and Analysis.

3.1 The flaw in the Objective.

Many major mobile brands like Apple, One-Plus and Samsung don't follow a linear popularity curve in which every subsequent model is better than the previous. This is based on the statistics provided by Statista (S.'Odea 2020).It is evident that Iphone 7 that was released in 2016 was the peoples favorite until 2020 approximately the same time this paper and work was under creation. This proves that comparing two companies over the popularity of a wide range of similar products is very generalized and cannot be used for a concrete analysis. Simply put, one might consider giving positive feedback for an outdated Samsung model in comparison to a much more advanced iPhone model, creating an unfair playing field. Hence, the objective of comparing two giant brands on a generalized forum isn't the best way to advance in this study.

3.2 Methodology Improvisation.

The pre-processing techniques used are very updated when it comes to the use of the Bag-Of-Words method, or the TF-IDF Model. However, there are some major setbacks that are enlisted below:

- The team has used Tokenizer to tokenize the data and not Twitter-Tokenizer provided by the NLP as nltk.tokenize (*nltk.tokenize package_2022* 2022) package which is mainly build for twitter tokenization.
- The use of min_df has made the model inertly robust to the frequency of the use of Keywords. This is a major setback, as the words used to boost or criticize the subject under consideration aren't limited, there is a high possibility that many individual words are not actually considered due to lack of frequency. For instance, out of 1000 tweets "brilliant" is a word used just 2 times, due to the min_df set to 2 this particular word will never be considered. Hence, this can cause huge fluctuations in the result.
- Finally, many tweets are expressed in emojis today, and there is no consideration of emojis or sarcasm encapsulated. This consideration would make the project more flexible and accurate.

3.3 Result Analysis

The analysis of the result is going to be pin-pointed on the Sanity check between the assumption of the result generated and the real world scenario. Primarily, as mentioned in the Objective Flaw, the comparison of all models in general cannot be a fit choice for the success of a company over the other. Secondly, there is no mention of any statistic report or study of the success amidst Samsung and IPhone rather than reviews on an unnamed online shopping site, which again isn't concrete to conclude the result. And Lastly, the observations in the Methodology used with min_df to gain a high accuracy can also be the cause of a really weak recall. A confusion matrix projection would be the right way to analyze the outcome, however it isn't provided.

4 Conclusion & Significance.

The work of the Team is commendable in many aspects considering the vision and the methods used. However, it lacks behind by using conventional methods and thus causes many inaccuracies in the claim of accuracy and statistical result deployment. The constant advancement in NLP is astonishingly portrayed by the timeline as many techniques used were the leading advancement in 2019 and now is redundant. I believe my insights and suggestions would definitely help this work in making it more fruitful and accurate.

References

Hasan, Md Rakibul, Maisha Maliha, and M Arifuzzaman (2019). "Sentiment analysis with NLP on Twitter data". In: *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*. IEEE, pp. 1–4.

nltk.tokenize package_2022 (2022). `https://www.nltk.org/api/nltk.tokenize.html`. [Accessed 07-Jun-2022].

S.'Odea (2020). *iPhone model share worldwide 2020 — Statista — statista.com.* `https://www.statista.com/statistics/626631/smartphone-market-share-by-device-worldwide/`. [Accessed 07-Jun-2022].