Review of: "Strategic Citations in Patents: Analysis Using Machine Learning"

Bhaskar Mukherjee[1]

1 Banaras Hindu University

Potential competing interests: No potential competing interests to declare.

The study has been conducted through a unsupervised ML algorithm and use cosine similarity for measuring the proximity of ideas. The dataset is taken from USPTO bulk data and Doc2Vec algorithm has been adopted for discussing the results.

The concept of the paper is quite good but very less of results and more with discussion. I think it would be better to explain the results in more.

This type of study is only possible with those dataset where the citation to the patents are available. As USPTO patents are not accessible freely, repetition of results was an impossible task for me.

What I feel is in scientific communication, the word 'I' is less in use. Either in common noun of collective noun is better option.

Although, the paper is explained that the '**bibliographic text from 1976 to end of 2015**",but I am unable to found the actual figure of the dataset on the basis of which author explained that "While 6.3% of target patents are directly cited when their similarity ranges between 0.5-0.6, only 4.2% are cited for similarity 0.6+." Is it "2,306,041 patents"?

As the author explained "due to strategic motives after the inventor changes firms".. is it mean that the study is based on Industrial patents only, or the inventors who developed industrial patents. It will be hopefully useful, if author can explain a bit about the characteristics of the data.

In my experience, when we lemmatize the scholarly word into its base/root word, sometime the word does not reveal the real meaning as it is revealed in the text. Stop words are very domain-specific. It would be better if author could enlighten a few in this regards.