International Journal of Strategic Engineering
Volume 5 • Issue 1

Comparing the Behaviour of Two Topic-Modelling Algorithms in COVID-19 Vaccination Tweets:

LDA vs. LSA

Jordan Thomas Bignell, Coventry University, UK*
Georgios Chantziplakis, Coventry University, UK
Alireza Daneshkhah, Coventry University, UK

ABSTRACT

Coronavirus is a newly developed infectious disease that has triggered a pandemic due to its ease of transmission as of early 2020. Several groups from various countries have been working on a vaccine to prevent and avoid the spread of the virus in this outbreak. In this article, the main objective is to compare LDA against LSA to gain a better understanding of the Tweets and which topic modelling technique fits best for this task, and additionally if the feedback of the Tweets were positive or negative sentiment. It was concluded that LDA was a better unsupervised technique for categorizing the raw text in 12 topics.

KEYWORDS

COVID, Hierarchical Dirichlet Process, Latent Dirichlet Allocation, Latent Semantic Analysis, Social Media, Twitter, Unsupervised Machine Learning, Vaccines

INTRODUCTION

As of early 2020, Coronavirus is a newly formed infectious disease and due to the ease of transmission has caused a pandemic. The official medical name for Coronavirus is COVD-19, which will be referred to throughout this paper (World Health Organisation, 2020). This pandemic has forced the majority of the world to change and adapt to a new lifestyle to avoid the spread of infection of this newly formed disease. It is found that older people and those with underlying health conditions such as diabetes and cardiovascular disease have a higher chance to develop more impactful symptoms that could be life-threatening. (World Health Organisation, 2020). Throughout this epidemic, many organizations from different countries have been developing a vaccine that can prevent and stop the spread of this virus. This would allow the public to return to normality, however, the vaccine cannot be developed and produced instantly due to the testing and rigorous experiments to ensure the safety of the vaccine. Due to the fierce and rapid development of the vaccine, many people have concerns. Therefore, obtaining the tweets to decipher and categorise each topic will provide an understanding of the views, and opinions of the vaccine, if this is positive or negative feedback.

DOI: 10.4018/IJoSE.292445

*Corresponding Author

Topic modelling is an unsupervised machine learning approach used for detecting abstract topics contained in a collection of text documents. The two main topic models used are Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). These techniques can detect words and phrase patterns, scanning documents, and automatically clustering word groups and similar expressions (Pascual, 2019). Some practical examples of topic modelling applications can be found in the papers by Ni Ki et al. (2021) and Daneshkhah et al. (2020), that discusses the applications of topic modelling in medicine, specifically diabetes, and effective preventive techniques for predicting behaviours linked to cybercrime, respectively.

Machine learning techniques have become more popular since the emergence of the COVID-19 pandemic, as been seen in the publication by Vepa et al. (2021), that introduces some methods for clinical decision-making tools for COVID-19 inpatients. The addressed problem of investigating COVID-19 further can be dealt with by analysing social media platform like Twitter. Moreover, the issue lies within the limited number of words that Twitter allows each user to post. Each tweet allows a maximum of 280 characters, this size potentially causes issues with the proposed algorithms, as they are more suitable for larger text groups. To counter the issue, the aim is to preserve only key words that are related to the subject, allowing for further investigation to be executed, inspecting the accuracy of topic modelling with the methods presented.

This paper aims to study and investigate the behaviour of LDA and LSA, on COVID-19 Tweets which were published between December 2020 and April 2021. This comparative study will provide beneficial information for small text groups, in addition, understanding the topics categorised by the tweets. The expectation of this paper is to learn the subject areas from each vaccine that has been included in this dataset. Furthermore, taking into consideration the social, ethical, and legal context of COVID-19. The aim of the paper is to investigate if LDA is a suitable topic modelling technique against LSA for small text groups.

LITERATURE REVIEW

In this publication by Garcia & Berton (2021), the paper was based on tweets in both English and Portuguese language, which from research studies have only been conducted in one language. The modelling techniques used in this paper were Latent Dirichlet Allocation (LDA) and Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) which is better suited for short text due to the lack of word co-occurrences.

For Sentiment Analysis specifically, CrystalFeel and SBERT are used to capture the range of emotions used in the text such as joy, sadness, anger, or fear. Due to the language difference, it is worth noting the change of keywords of certain subject areas such as ‘pandemic’ in English and ‘Pandemia’ in Portuguese, and ‘Quarantine’ and ‘Quarentena’ respectively. Therefore, teaching the system to understand these terms is important to learn and correctly categorise each topic. Alternatively, creating 2 separate systems to work dependable on language could provide better results. In total there was a recorded 7,144,349 English tweets and 7,125,530 Portuguese tweets, furthermore, conducting the machine learning techniques on language could output more defined results given the large dataset for each language, respectively.

From the results section of the paper (Table 6 & 7, Garcia, K., & Berton, L. 2021), it is clear from the 10 selected topics of Economic impact, Case reports, Proliferation care, politics, entertainment, treatments online events, charity, sports, and anti-racism protests, that the Portuguese topics have been categorised using more words in comparison to the English topics for example:

- English topics for Economic impact: Work, impact, business, crisis, pay.
- Portuguese topics for Economic impact: Buying, money, working, crisis.