Trends in COVID-19 Publications: Streamlining Research Using NLP and LDA

Akash Gupta¹, Shrey Aeron², Anjali Agrawal³ and Himanshu Gupta⁴* 

¹ Department of Engineering, University of Cambridge, Cambridge, United Kingdom, ² Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, United States, ³ Harmony School of Innovation – Sugar Land (High School), Sugar Land, TX, United States, ⁴ Valley Health System, Ridgewood, NJ, United States

Background: Research publications related to the novel coronavirus disease COVID-19 are rapidly increasing. However, current online literature hubs, even with artificial intelligence, are limited in identifying the complexity of COVID-19 research topics. We developed a comprehensive Latent Dirichlet Allocation (LDA) model with 25 topics using natural language processing (NLP) techniques on PubMed® research articles about “COVID.” We propose a novel methodology to develop and visualise temporal trends, and improve existing online literature hubs.

Our results for temporal evolution demonstrate interesting trends, for example, the prominence of “Mental Health” and “Socioeconomic Impact” increased, “Genome Sequence” decreased, and “Epidemiology” remained relatively constant. Applying our methodology to LitCovid, a literature hub from the National Center for Biotechnology Information, we improved the breadth and depth of research topics by subdividing their pre-existing categories. Our topic model demonstrates that research on “masks” and “Personal Protective Equipment (PPE)” is skewed toward clinical applications with a lack of population-based epidemiological research.

Keywords: natural language processing, latent dirichlet allocation, COVID-19, trends, LitCovid, topic model, Pubmed

INTRODUCTION

The COVID-19 outbreak was officially declared a pandemic by the World Health Organization in March 2020 (1). As the number of COVID-19 cases and deaths has increased, so has the research. Searching “COVID” in PubMed®’s database gives a list of over 32,000 unfiltered publications (as of July 2020). Due to the overwhelming stream of papers, there is now an urgent need for tools to automate the categorical organisation of research. More importantly, to sufficiently address COVID-19 and future pandemics, it is necessary to streamline the research and development process by allowing for quick identification of research areas that are either gaining popularity or lacking adequate research.

Latent Dirichlet Allocation (LDA) is an unsupervised topic modelling technique used to learn hidden topics within a corpus (2). It assumes topics are a soft clustering of words and outputs two probability distributions: a distribution of topics in the corpus, and distributions of words across each topic. Currently, LDA, with the aid of natural language processing (NLP) methodologies, has been used to investigate the response to government policies (3), analyse public sentiment on social media (4) and the news (5), and understand general
research hotspots in publications (6) as well as global trends (7). Most of these topic models either use a small number of topics or group a large number of topics into overarching themes, which eases comprehension at first glance. However, for a more in-depth analysis, a better understanding of the complexity of topics within each theme is required, which cannot be captured by searching for that topic in a literature repository using a simple query.

The National Center for Biotechnology Information (NCBI) has developed LitCovid, a central repository for curated COVID-19 research (8). The hub uses a combination of human and machine-learning methods to provide

![Frequency charts of abstracts published in PubMed®.](image)

**Figure 1** | Frequency charts of abstracts published in PubMed®. (A) Per day frequency shows a weekly pattern. (B) Per week frequency from January 17th to July 17th 2020 with a gaussian curve fitting (shown in red dashed line). The training set was split from the testing set at the 81% mark on June 22nd 2020. The weekly release pattern informed our decision to analyse trends on a per-week basis.
publications categorised by eight topics alongside temporal and geographic distributions. Although very useful, the defined categories are overly broad, limiting rapid assimilation of the ongoing research.

Here, we describe and implement an NLP methodology using LDA to identify topics in COVID-19 research. We then visualise and evaluate temporal trends to recognise the dominant and under-represented areas of research. We also apply our methodology to the LitCovid dataset and further subdivide their categories into topics. Although our results focus on COVID-19, we believe our method is generalisable and sustainable for evaluating rapidly evolving research fields.

**METHODS**

**Data Sourcing**

Our dataset is extracted using Entrez Programming Utilities (9) from the PubMed Application Programming Interface (API) through the BioPython package. The API returns a list of metadata on all documents retrieved in a search query, which was the term “COVID.” For each abstract, the PubMed Document ID (PMID), date of publication on PubMed, and abstract are stored.

A similar process is repeated for LitCovid abstracts. LitCovid (8), aided by machine-learning methods, manually assigns the articles into the eight categories of “Mechanism,” “Transmission,” “Diagnosis,” “Treatment,” “Prevention,” “Case Report,” “Forecasting,” and “General.” We use the NCBI Coronavirus API to extract the PMIDs corresponding to abstracts in each LitCovid category. We perform the same data cleansing process on the LitCovid abstracts as with the PubMed abstracts.

For significantly large corpus sizes with a variety of journals, abstracts produce very similar topics compared to those produced by using the full text (10). Hence, we decided to use abstracts (3) due to the further benefits of much-reduced computation and free accessibility, aiding in reproducibility. A small number of abstracts are excluded from the corpus ($N = 99$) for the following reasons: those with <50 characters ($N = 73$), and others with a sentence including “This corrects the article DOI [...]” ($N = 26$).

**Overview of the LDA Model**

LDA is a standard topic modelling algorithm that learns the hidden topic structure within a corpus (2). Each topic has a weight within the corpus and is represented by a set of related words. We used Gensim’s Python implementation of LDA to develop a topic model using a bag-of-words representation of the pre-processed training set. Further details about the NLP methodology and optimisation process are explained in the Supplementary Material. In brief, we set an upper limit on the number of topics to 50 and optimised hyperparameters based on the coherence value. Additionally, we selected a random seed.
to use in the training process to ensure replicable results. We implemented the methodology in Python.

**Temporal Evolution of Topics**

To evaluate the temporal trends, we propose a novel method, which is applied to both PubMed® and LitCovid abstracts to produce an intuitive visualisation of the weekly temporal evolution of topic proportions. Analysis on a day-by-day basis can result in too much noise; instead, the documents are grouped by week, which is further justified by the apparent weekly release pattern of papers as shown in Figure 1A. However, since the weeks in January 2020 contain a substantially low number of papers, they are grouped into a single month. To evaluate how well the temporal evolution can predict future releases, we assign the weeks into training and testing sets with an approximate ratio of 8:2 (which is shown by the test-train split in Figure 1B).

For each week, in both the training and testing sets, all abstracts are combined into a single document. Applying the LDA model to these documents returns an estimate of the proportions of each topic that week. These results are graphed to produce a temporal topic evolution that is compared between PubMed® and LitCovid. We also create a heatmap to visualise all topics in the corpus in a concise chart.

**Subdivision of LitCovid Categories**

For each of the eight LitCovid categories, all abstracts are combined into a single document. Applying our model to the documents returns the proportions of each LDA topic in each LitCovid category. Statistical analysis (as described in the next section) is used to evaluate the categories and subdivide them into corresponding LDA topics.

**Statistical Analysis**

We incorporate tools from the Gensim Package (11) to perform statistical analyses on the topic model. We use perplexity (how surprised a model is to a sample) and topic difference to track the model convergence. The final performance of the model is evaluated using a coherence score (12).

Since LDA is a probabilistic model (2), the methodology used to capture temporal trends, as well as the subdivision of LitCovid categories is statistical. Applying the LDA topic model to a document returns a proportion for each topic, which is then used for further analysis.

Due to the non-linear nature of the temporal trends, a qualitative account of the trend fitting between the testing and training sets is discussed. Temporal trends of PubMed® and LitCovid publications are compared using Normalized Euclidean Distance, and topics are compared to each other.

### Table 1: Topic description and summary about COVID-19 produced by LDA model.

| Topic | Top 10 Most Relevant Words | Topic summary |
|-------|-----------------------------|---------------|
| 1     | Care, pandemic, service, telemedicine, health, care, need, resource, challenge, telehealth, healthcare | Health care, Telemedicine |
| 2     | Number, case, model, country, estimate, epidemic, estimated, data, daily, rate | Epidemiology |
| 3     | Case, day, chest, pneumonia, lesion, symptom, consolidation, fever, group, showed | Pulmonary |
| 4     | China, outbreak, epidemic, world, health, prevention, control, case, disease, novel, coronavirus, public, health, january | Disease Outbreak |
| 5     | Review, article, pandemic, research, vaccine, current, paper, scientific, literature, evidence | Scientific Development |
| 6     | Surgery, surgical, procedure, surgeon, hospital, dental, ppe, pandemic, emergency, staff | Surgery |
| 7     | Mortality, study, compared, included, outcome, risk, factor, higher, group, hospitalized, severe | Clinical Outcomes |
| 8     | Participant, survey, anxiety, questionnaire, mental, health, respondent, psychological, stress, fear, perceived | Mental Health |
| 9     | Disease, cardiovascular, acute, respiratory, cardiac, distress, syndrome, severe, ards, cytokine, storm, syndrome, cardiovascular, disease | Cardiovascular |
| 10    | Treatment, trial, clinical, trial, hydroxychloroquine, drug, study, remdesivir, hcq, tocilizumab, therapy | Clinical Trial |
| 11    | Cell, ace2, expression, receptor, sarscov2, tissue, human, virus, lung, pathway | Pathogenic Mechanism |
| 12    | Test, assay, testing, detection, sample, sarscov2, positive, specimen, sensitivity, antibody | Diagnosis |
| 13    | Child, symptom, pediatric, neurological, report, infection, infant, case, sarscov2, reported | Paediatrics |
| 14    | Social, crisis, health, economic, pandemic, impact, policy, consequence, pscinfo_database, right, reserved | Socio-economic Impact |
| 15    | Cancer, icu, intensive, care, treatment, ventilation, mechanical, ventilation, requiring, therapy, lung, cancer, ecmo | Oncology |
| 16    | Recommendation, risk, guideline, consensus, healthcare, worker, guidance, management, ibd, transplant, expert | Guidelines |
| 17    | Protein, sarscov2, compound, drug, target, vaccine, spike, protein, binding, inhibitor, epitope | Virus Structure |
| 18    | Virus, sequence, genome, human, bat, sarscov2, mutation, coronaviruses, genetic, animal | Genomic Sequence |
| 19    | Resident, person, county, older, adult, household, state, united, state, black, population, among | Demographics |
| 20    | Level, coagulation, thrombosis, ddimer, severe, elevated, crp, platelet, coagulopathy, aki | Haematology |
| 21    | Social, medium, information, public, video, news, medium, tweet, hand, misinformation, india | Communication |
| 22    | Use, vitamin, medication, drug, arb, angiotsin, angiotsin, receptor, ace, inhibitor, blocker, reninangiotensin, system | Existing Treatments |
| 23    | Liver, respiratory, skin, coinfection, ultrasound, gastrointestinal, symptom, diarrhea, imaging, symptom, fecal | Gastroenterology |
| 24    | Mask, aerosol, device, droplet, blood, particle, app, airborne, filter, mobile | Airborne transmission protection |
| 25    | Pregnant, woman, pregnancy, woman, maternal, pregnant, birth, neonatal, delivery, suspected, radiology, department | Pregnancy |

After optimising the LDA model, 25 topics were produced. The topics are listed in descending order of prevalence in the corpus. The top 10 words (middle column) are chosen based on relevance to the topic, with lambda = 0.6 (13). The topic summary (last column) is based on the consensus of the authors after reviewing the top 20 words and the top 10 abstracts.
using Jaccard Distance, a comparison between disjoint terms in two models (13).

LitCovid categories are compared against each other as well as the PubMed® corpus using Hellinger Distance (H):

\[ H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum \limits_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2} \]

where \( k \) is the number of topics; \( P \) and \( Q \) are the topic proportions returned by applying our LDA model to each LitCovid category.

**RESULTS**

The PubMed® API returned 16,445 abstracts with the search query “COVID” (as of July 17th, 2020). We exclude 73 abstracts with <50 characters and 26 abstracts with extraneous information, resulting in 16,346 abstracts in the corpus from January 17th to July 17th, 2020. The distribution of these follows a gaussian curve with mean on June 9th, 2020 (Figure 1B). The corpus is split (81:19) as per section Temporal Evolution of Topics: 13,212 abstracts for training and 3,134 abstracts for testing. The cutoff date for the training set is June 22nd (inclusive).

LitCovid generates 27,595 abstracts (as of September 8th, 2020) from the entirety of their dataset. We exclude 13 empty abstracts, 97 abstracts with <50 characters, and 37 abstracts with extraneous information. The final number of abstracts is 27,448, ranging from January 17th to September 5th 2020. 4,223 LitCovid abstracts could not be associated with any category and 23,225 are associated with at least one of the eight categories. The distribution of these categories is in Figure 2.

The LDA model is trained on 13,212 abstracts to produce a topic model with 25 topics. The final perplexity is 185.6, and the coherence score is 0.526. The average normalised Jaccard distance between the topics is 0.923 (STD 0.048) with a minimum of 0.701 between Topic 23 ("Gastroenterology") and Topic 13 ("Paediatrics"). A list of all the topics is displayed in Table 1 alongside the top 10 relevant words and our summary interpretation based on the top words and abstracts associated with each topic. The topic of “Health care, telemedicine” has the highest proportion of 9.4%, whereas “pregnancy” has the lowest proportion of 1.0%.

Figure 3 provides a general breakdown of how topic proportion changes over the timeframe. The weekly topic proportions for the PubMed® test data were all within the bounds of the training data and are plotted as the top three rows in the heatmap, which show hardly any visible discrepancies. Most topics start at some high/low proportion, then trend in the opposite direction and begin to level off in April 2020. Others follow a relatively steady change with time. The highest topic proportion of 41% is during January 2020 in Topic 18, “genomic sequence.” Topic 4, “disease outbreak,” has the highest mean weekly topic proportion of 13.5% (STD 10.5). Topic 24, “airborne transmission protection,” has the lowest mean of 0.98% (STD 0.66). This relatively lower topic proportion, as also illustrated in Figure 4A, was striking, because the usage of masks is a
highly emphasized guideline (16). We performed a post hoc analysis by analysing the proportion of the term “mask” (as shown in Figure 4B) in different topics: Topic 24 ("airborne transmission protection") 2.3%, Topic 16 ("guidelines") 0.078%, Topic 6 ("surgery") 0.033%, Topic 2 ("epidemiology") 0.0090%, Topic 8 ("mental Health") 0.018%.
Applying our LDA model to LitCovid abstracts shows that temporal topic proportions have a very strong agreement with an average Normalized Euclidean Distance (NED) of 0.0303 (STD 0.0128). The most dissimilar looking graph is shown in Figure 5A, which is Topic 17, “Virus Structure” with NED 0.0604. Even though LitCovid’s dataset was also taken from PubMed®️, the mild variation can be attributed to differences in search queries (as described in Figure 2). The most similar trends occur for Topic 1 (“Health care, telemedicine”) with NED = 0.0303, as shown in Figure 5B.

In Figure 6, applying our model to the LitCovid categories shows that the LitCovid category “Epidemic forecasting” is most similar to LDA Topic 2, “Epidemiology” as it has a topic proportion of 74%. The most intuitive category to subdivide is “General Info”; our topic model split it into “Scientific Development” (21%), “Health Care, Telemedicine” (16%), “Socio-economic Impact” (15%), “Disease Outbreak” (14%), “Genomic Sequence” (9%) and “Communication” (6%). Another notable subdivision is of the “Case Report” category into the relevant medical fields of “Paediatrics” (20%), “Pulmonary” (18%), “Oncology” (13%), and “Cardiovascular” (12%), as well as identifying the under-represented topics “Haematology” (5%) and “Gastroenterology” (6%).

As shown in Figure 7, the LitCovid categories with the most overlap are “Treatment” and “Mechanism” with H = 0.160, whereas the least overlap is for “Epidemic forecasting” and “Case Report” with H = 0.572. When compared to the PubMed®️ corpus, “Prevention” has the lowest Hellinger distance of H = 0.116, suggesting that the “Prevention” category may be too broad. Conversely, “Epidemic Forecasting” has the greatest Hellinger distance (H = 0.476), suggesting that it is a well-defined category.

**DISCUSSION**

We provide a generalisable NLP methodology to extract abstracts from PubMed®️, create an optimised LDA topic model, and visualise temporal trends. Applying our model to LitCovid abstracts helps to identify trending and under-represented research areas as well as subdivide its categories into relevant
topics. We find that an important topic related to masks and PPE is under-represented in population-based research.

The optimised LDA model developed in this study identifies 25 topics with no significant overlap, and all relate to the COVID-19 pandemic. Our model is trained and tested on a much larger number of documents compared to previous studies, and we believe our topic interpretation is improved using NLP techniques such as the use of bi-grams and ordering terms by relevance rather than frequency (further details are in Supplementary Material). Therefore, our model is comprehensive, and our analysis is in-depth. Topics not previously identified include “Pathogenic Mechanism,”
“Cardiovascular,” and "Demographics" (Topics 9, 11, and 19). Previous topic models developed using research publications provide broad categories and do not sufficiently characterise the social science aspect of COVID-19 (6, 7, 17, 18). In contrast, our model covers this field in the topics “Mental Health,” “Socio-economic Impact,” and “Communication” (Topics 8, 14, and 21, respectively), representing 10.2% of the PubMed® test corpus.

We developed a customised heatmap shown in Figure 3 representing topic trends on a holistic scale. This visualisation allows for intuitive understanding compared to the use of classic line graphs, which can be cumbersome to review, especially for a larger number of topics. From January to March 2020, Topic 18, “genomic sequence” has an unusually high topic proportion given its lower proportion in the overall corpus. This trend correlates well with our expectation as it was initially a hot topic in social media (19). The heatmap also clearly shows that Topic 24 (“Airborne Transmission Protection”) has a lack of adequate representation. As a direct application of our topic model, we zoomed into the word “mask” (Figure 4B), demonstrating its representation in very few topics, including Topic 24, 16 (“guidelines”), and 6 (“surgery”). Moreover, the relative weight of “mask” within Topic 16 is two orders of magnitude less (“guidelines”), and 6 (“surgery”).

Figure 7 shows there is measurable overlap between LitCovid’s categories, some of which is to be expected because they all contain the overarching theme of COVID-19. Interestingly, our model shows that the most heterogeneous category, “Prevention,” comprises the topics “Health Care,” “Telemedicine,” "Epidemiology,” ”Surgical Procedures,” and “Mental Health.” One would not expect some of these, but rather it would be more reasonable to have “Guidelines” as a core topic. The closest categories identified in Figure 7 are “Treatment” and “Mechanism,” which is reasonable because effective preventative treatment should be inhibiting a mechanism. Closer analysis using Figure 6 shows that “Mechanism” is almost a subset of “Treatment,” with the key difference being that “genomic sequence” is a strong topic in “Mechanism,” but not in “Treatment.” We believe that these categories can be ambiguous to a researcher, and our methodology would significantly improve online literature hubs through more specific subdivisions.

As we performed searches in LitCovid, we noticed that abstracts were sometimes not relevant to the query even though a relevance option is provided. Improved comparative relevance of a topic within a research paper can add value for researchers as it distils a large mass of research papers by the strength of their major topics. Future research can utilise our methodology to enhance online literature hubs by ordering documents based on the proportion of topics.

Although the LDA topic model can be updated with new research, further investigation on hyperparameter adjustment will be required to identify new topics. However, our methodology is simple to re-run and provide up-to-date trends. To evaluate the temporal trends, we use the date of publication, because we found that not all articles had the research date. Although the publication date does not accurately reflect the time when the research was performed, for our purpose, our temporal visualisation disseminates the trending topics from the under-represented ones, which we believe is more crucial to the researchers. Another potential limitation in our dataset is the exclusive use of English abstracts. Since LDA is essentially a clustering algorithm, if more than one language is used, the topic model will likely return duplicate topics in different languages, which is redundant for our purpose. Instead, since many research papers written in other languages provide abstracts translated into English (22), this further justifies our use of abstracts instead of papers. Even though a significant number of publications are not analysed due their lack of abstracts, we believe this would not affect our identification of topics, given that it is reasonable to assume that this is random and not biased against any topic.

NLP techniques applied with LDA topic modelling results in a comprehensive topic listing and identification of important temporal trends. Our methodology has the potential to complement existing literature hubs. We identify topics for further research, such as studies on masks that may be of significant public interest.
DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SA, AA, and AG: data generation. AA, SA, AG, and HG: data analysis and manuscript. AG, SA, and HG: statistical analysis. HG: final manuscript approval. All authors contributed to the article and approved the submitted version.

REFERENCES

1. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Biomed. (2020) 91:157–60. doi: 10.23750/abm.v91i1.9397
2. Campbell JC, Hindle A, Strouli E. Latent dirichlet allocation. The Art and Science of Analyzing Software Data. Elsevier (2015). p. 139–59.
3. Debnath R, Bardhan R. India nudges to contain COVID-19 pandemic: a reactive public policy analysis using machine-learning based topic modelling. PLoS ONE. (2020) 5:e0238972. doi: 10.1371/journal.pone.0238972
4. Orud C, Purushotham S, Raff E. Exploratory Analysis of COVID-19 Tweets using Topic Modeling, UMAP, and DiGraphs. (2020). Available online at: http://arxiv.org/abs/2005.03082 (accessed September 15, 2020).
5. Liu Q, https://pubmed.ncbi.nlm.nih.gov/?term=Zheng%5BSAuthor%5D&cauthor_id=32302966Chen Q, Liu G, https://pubmed.ncbi.nlm.nih.gov/?term=Chen%5BS%5D&cauthor_id=32302966Chen S, et al. Health communication through news media during the early stage of the COVID-19 outbreak in China: digital topic modeling approach. J Med Internet Res. (2020) 22:e19118. doi: 10.2196/ 19118
6. Dong M, Cao X, Liang M, Li L, Liu G, Liang H. Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modeling. medRxiv. (2020). doi: 10.1101/2020.03.26.20044164
7. Tran BX, https://www.ncbi.nlm.nih.gov/pubmed/?term=Ha%20GH%5BSAuthor%5D&cauthor=true&cauthor_id=32517764aGH Nguyen LH, https://www.ncbi.nlm.nih.gov/pubmed/?term=Vu%20GT%5BSAuthor%5D&cauthor=true&cauthor_id=32521776VuGT, Hoang MT, https://www.ncbi.nlm.nih.gov/pubmed/?term=L%5BSAuthor%5D&cauthor=true&cauthor_id=32521776LeHT, et al. Studies of novel coronavirus disease 19 (COVID-19) pandemic: a global analysis of literature. Int J Environ Res Public Health. (2020) 17:4095. doi: 10.3390/ijerph17114095
8. Chen Q, Allot A, Li Z. Keep up with the latest coronavirus research. Nature. (2020) 579:193. doi: 10.1038/d41586-020-00694-1
9. Entrez Programming Utilities (E-Utilities). Encyclopedia of Genetics, Genomics, Proteomics and Informatics. Dordrecht: Springer (2008). p. 612.
10. Syed S, Spruit M. Full-Text or Abstract? Examining topic coherence scores using latent dirichlet allocation. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA). Tokyo (2017).
11. Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. (2010). Available online at: http://citeeseerx.ist.psu.edu/viewdoc/ summary?doi=10.1.1.695.4595 (accessed October 6, 2020)
12. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15. Shanghai (2015).

ACKNOWLEDGMENTS

We would like to thank Aditya Gupta, MEng, for his valuable suggestions related to machine learning, the Entrez API, and topic classification. Our research paper appears on a preprint server from Social Science Research Network (SSRN) (23).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth. 2021.686720/full#supplementary-material

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Gupta, Aeron, Agrawal and Gupta. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.