Topic Modelling for Inter7s Marketing Development

(May 2024)

Bolormaa Mendbayar – x23176725
Thapelo Khantsi – x23131535
Temitope Oladimeji – x23187204
Ziyi Yan – x22198512

National College of Ireland
MSc in Data Analytics
(MSCDAD_A)
Data Mining & Machine Learning 2

***Abstract* — This report delves into a novel application of topic modeling to improve marketing strategies for Inter7's, exploring LDA to identify potential business partnerships within the mental health and sports sectors. Through the application of LDA, textual data was analyzed from 200 companies to discover hidden thematic patterns and assess thematic alignment between companies. Using both bar charts and interactive pyLDAvis visualizations, the analysis offered insights into the prevalent themes and their distribution across the data. Additionally, an analysis of topics and their associated keywords was conducted to uncover relevance within the corpus of documents. This process revealed the most dominant topics in the context of marketing strategies. Through this evaluation, the study contributes to advancing the understanding and application of topic modeling techniques in marketing strategy development, particularly in the mental health and sports sectors. The framework implemented also allows for more effective use of topic modeling methodologies in various domains.**

Keywords — Latent Dirichlet Allocation (LDA), topic modelling, marketing strategy, mental health, business partnerships

I. Introduction

In today's market, companies are constantly seeking unique strategies to improve their marketing efforts and gain a competitive advantage. While traditional methods have been successful over the years, the rise of artificial intelligence has paved the way for even more possibilities, with an increasing demand for innovation. The abundance and complexity of data has given companies the opportunity to implement these advanced technologies, with the goal of providing insights specifically tailored to their needs.

This project explores a unique application of topic modelling to improve marketing strategies for Inter7's. Instead of utilizing topic modelling solely for text classification or clustering analysis, it was used differently to identify hidden patterns within textual data from over 200 companies. The project seeks to answer the following research question: How can topic modelling, specifically Latent Dirichlet Allocation (LDA), be used as a marketing technique to identify potential business partnerships in the charity, mental health, and sports sectors?

The significance and novelty of this research lies in its approach to using topic modelling as a tool for marketing purposes. By extracting relevant topics and keywords from company descriptions, the aim was to find the companies associated the most with these topics, subsequently identifying potential partnership opportunities.

The objectives of this project were as follows:

- To apply LDA to analyse textual data from over 150 companies and identify the most prominent topics based on a given criteria.
- To extract relevant keywords associated with each topic to gain deeper insights into the focus area of each company.
- To determine the optimal company and its associated topic probability within the charity, mental health, and sports sectors.
- To recommend potential business partnerships by identifying companies with a high topic probability (> 0.9) aligned with the company's interests.

While these objectives may seem ambitious, they represent a good opportunity to explore other aspects of topic modelling, especially in marketing development. However, challenges may arise in interpreting the results, ensuring the relevance of identified topics and keywords, and therefore translating these into strategies.

The project aims to demonstrate the versatility of topic modelling, and its applicability in strategic planning. The following sections explore this approach in more detail:

a) Related Work: A review of existing literature and research in the area to validate the project's findings.

b) Data Mining Methodology: An outline of the specific steps taken to collect, preprocess, and analyse textual data, including the application of LDA.

c) Evaluation/Results: Presentation of the analysis and findings, including the identification of prominent topics and recommendations.

d) Conclusions and Future Work: Summary of key insights, and discussion of limitations and areas for future research.

II. Discussion Of Related Work

Marketing strategies were the key focus in [1], which aimed to offer insights into the effectiveness of different topic modelling approaches including Latent Dirichlet

Allocation (LDA) and Latent Semantic Analysis (LSA), from both a technical and marketing point of view.

The study's methodology involved the classification of extracted topics into distinct categories. This was done by implementing keywords to be used during analysis, specifically when preprocessing the document. A comparative assessment of each topic modelling technique was performed, highlighting LDA as the most effective. The categorization and technique comparisons were some of the study's biggest strengths as it researches about the best technique during practice. There is also practical relevance from the study's conclusion for data analysts and marketing researchers. However, despite the advantages, the study's corpus size may limit the generalizability of its findings. Also, the manual classification of extracted topics into categories may introduce bias into the analysis. Automated techniques can ensure validity in the results.

Alternative techniques such as BERTopic and Top2Vec were recommended in the study. These can provide novel insights on the topic in the paper.

LDA has become a common approach to topic modelling [2], primarily due to its easy implementation compared to other techniques. In an education-focused study in [3], LDA was used to analyse students' responses to very broad questions. While the context differs from the idea of marketing applications, the principles of applying LDA remain consistent as it revealed relevant insights from textual data. In reference to the research question which centres on identifying companies with a strong commitment to mental health policies, the challenge of interpreting textual documents was also encountered.

The recommendation for LDA as a topic modelling approach was also echoed in [4], which analysed information extracted from social media. There is now an increasing use in marketing fields, as the extraction of key topics can help businesses identify potential leads and allow companies to reach out to other relevant organizations that share similar values. This aspect is a major discussion in [5], which focused primarily on how NLP can help in sentiment analysis. Analysing sentiments expressed in company descriptions within the charity, mental health, and sports sectors not only reveals topic relevance, but also a company's stance within their respective industries. The findings in [5] revealed that the model was accurate in its classification 96% of the time, also identifying the most significant contributing factors using a combination of deep learning and reinforcement learning techniques. The integration of sentiment analysis decomposition of LDA, which is a probabilistic method, can always offer deeper insights.

A major application of topic modelling is the analysis of co-occurrence patterns of words across documents. In [6], this approach was taken to separate these documents into six distinct topic, extracting relevant keywords from each. The study aimed to discover trends in the research within the field of cryptography, with its findings revealing an even distribution of papers across different topics. Topic modelling typically assumes a mixture of topics is contained in each document, with each topic being characterized by a distribution of words.

The automatic detection of main themes within these texts was discussed in [7]. To develop a potentially deployable system, the project focused on collecting large volumes of textual data, which LDA works well on. Improving the quality and accuracy of the process involved an extensive preprocessing stage, including tokenization, stemming and stop-word removal, which was also used in this research. Emphasis was placed on its novelty as well, advancing text analysis techniques in discovering main text themes.

Topic modelling has seen novelty in its use by researchers. Sentiment analysis and information retrieval are common cases where the model has been employed. However, recent studies have applied it to discover recurring themes within a specialized corpus. An example of this implementation is in [8], which applied topic modelling to English manuals for Psychological First Aid (PFA). Though the study acknowledged topic correlations with their respective document, the drawback to using LDA was the inability to perform a progressive analysis. Another drawback was the bias found during analysis, which limits its generalizability.

Bias in topic modelling can occur in many ways. Language bias is common because algorithms trained on a data from a specific language may inaccurately capture similar topics in other languages. This bias was introduced in [9]'s work on topic modelling using insights from COVID-19 where the data was limited to English tweets, overlooking insights from tweets in Indian languages. As noted in the study, even with India's spoken diversity topic, a biased representation of the virus's discourse in India was still unavoidable. Also, regional expressions may not make sense in other languages, adding to the challenge of cultural and language bias.

The choice of hyperparameters might also introduce bias during analysis, as observed in [10]. In this study, optimal hyperparameters for the LDA model were selected, which gave accurate results. However, hyperparameter tuning can result in skewed result if not performed carefully. Firstly, the study tested a limited number of hyperparameter combinations to obtain a perplexity score used to measure the model's generalization, potentially overlooking alternative configurations that could yield better-performing models. The narrow exploration of hyperparameter space may favour certain parameters over others, subsequently introducing bias in the findings. Also, the model's algorithm might be sensitive to initialization due to its manual selection of starting value. Though hyperparameter tuning is an essential part of any preprocessing stage for model analysis, algorithmic bias can easily be introduced, specifically affecting reproducibility.

A parallel to the research on mental-health related topics can be seen in [11], which investigated topic within text messages for mental disorders. The study used topic modelling techniques such as Latent Feature Dirichlet Multinomial Mixture (LFDMM) and Word Network Topic Model (WNTM). Similarly, the study in question specifically used LDA to discover topics within textual description of companies' policies regarding mental health, charity and sports. One advantage of the paper's approach is its use of multiple techniques, offering a detailed understanding of the effectiveness of different approaches. By assessing the performance of the techniques using different metrics, the study provides additional strengths and limitations for each of them.

However, one limitation of the paper's methodology is its reliance on a small dataset. While the study acknowledges the challenges in accessing large datasets of text messages due to privacy concerns, the limited sample size may affect the finding's generalizability. A larger and more diverse dataset would improve the result's validity. The paper's evaluation of the techniques is largely based on quantitative metrics, without considering the qualitative metrics such as the interpretability of the extracted topics, especially regarding topic coherence and relevance.

III. Data Mining Methodology

This This project follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology which entailed the following steps:

- **Business Understanding**: This stage involved understanding what the company needs, and the problem statement was created from the understanding to set the goals and objectives.
- **Data Understanding**: This involved understanding what data the company has, in order to inform the decisions of the machine learning and deep learning models to be considered for the study.
- **Data Preparation**: This is where gathered data was pre-processed, and organized to fit the modelling purposes.
- **Modelling**: The selected model was applied at this stage. Test designs and model assessments were pre-defined to evaluate the model.
- **Evaluation**: Based on the objective, the model was assessed to provide a business solution and recommendations to the company. The process of modelling was reviewed at this stage and results were validated based on the business objective set.
- **Deployment**: The stage entails details of how the stakeholders would access the results. Based on the goal, the final phase was the task of producing a final report that had a summary of the details of the project and the data mining results for actionable insights.

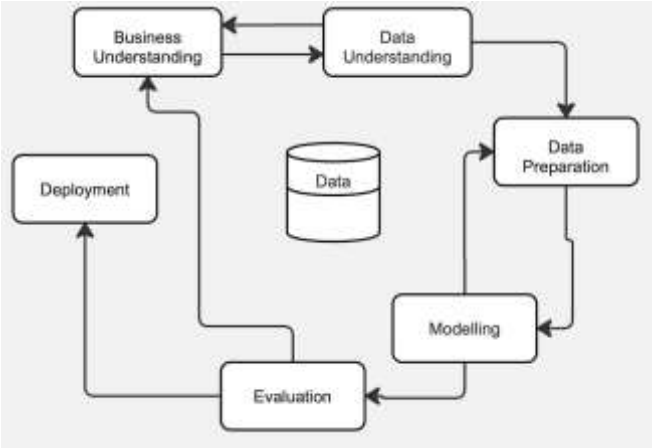


Figure 1: CRISP-DM

A. Business Understanding

The company's objective is to harness the benefits of AI to expand its leagues outside Dublin for growing their Dublin leagues and expanding their funding to mental health initiatives. Additionally, refining their marketing efforts to support league expansion.

B. Data Understanding

Based on the business objective. The understanding of data was a logical step followed next to understand how the company's data can be harnessed and mined to create a project that would be useful and provide actionable insights. The company's data was not sufficient nor useful in creating any machine learning or deep learning models for tackling the problem statement. Therefore, data was created instead. One crucial objective for data mining as outlined in [12], is to identify relevant data that will significantly impact the state of the art of the modelling and predictions being carried out, hence the significance of the stage in the project. Data creation was through the process of manual web scrapping, using predefined keywords for searching relevant data sources that are significant to our business objective. Which is finding data on companies that have similar concepts to Inter 7s, or companies in the space of mental health and sports which could be potential partners for the business marketing strategies.

1) Challenges of Web Scrapping

Web scraping is an automated way of extracting or harvesting data from the internet. Manual data harvesting was followed due to challenges regarding legal considerations. Many websites prohibit web scraping and doing so could be a violation of the terms and contract of the site. The EU database directives provide database protection through copyright therefore extracting data in this way would lead to an infringement. The application of intellectual property laws in regard to web scrapping remains uncertain in Ireland, even though there are cases where website terms and conditions have been upheld in court [13]. As a result, it was important to comply with the legal entities and take caution with manually extracting textual data from the internet based on the project's needs.

C. Data Preparation

Based on the business objective. The understanding of data was a logical step followed next to understand how the company's data can be harnessed and mined to create a project that would be useful and provide actionable insights. The company's data was not sufficient nor useful in creating any machine-learning models.

1) Data Cleaning and Pre-processing

The textual data harvested from the web had a lot of inconsistency and characters in it, making it hard to work with hence the necessity for this procedure. In the process, text cleaning and normalization was performed such as converting text to lowercase, removing punctuation and lambda function to replace any occurrence of commas, period, question marks and exclamation marks ensuring consistency in words presentation [14].

Performed future engineering on textual data by implementing the following with the help of natural language processing functions:

- **Tokenization**: By splitting the text words using the python built-in 'split' function into individual words or tokens. In this step, short words are removed from textual data and performed lemmatization to reduce words to their root forms. This helps to make sure that the data is standardized for analysis and assuring the effectiveness of the model by reducing the noise and irrelevant information.
- **Phrase Modelling (Bi-grams)**: This was performed using Genism's phrases which autom atically identifies and captures frequently occurring phrases within a text collection (corpus). This was important for enhancing the representation of text data for topic modelling and preserving the semantic meaning [15].
- **Document-Term Matrix(DTM) Creation:** Since the data used in qualitative, the DTM converts the data into numerical formats making it suitable for topic modelling. With this, comparison across documents can be achieved.

The number of topics provided for training was initially inputted manually which led to the "Goldilocks" problem; which specifies that if there are few topics, the important themes can be missed and if there are too many topics, we would end up with an overly granular and potentially meaningless breakdown of the data [16]. As a result, the optimal number of topics had to be determined using the coherence score technique, which is one of the few techniques that automatically determines the optimal number of topics. This was done utilizing Gensim's 'LdaMultiscore' which computes the score using the 'CoherenceModel' class from Gensim. The automatic detection of an optimal number of topics eliminates the need to manually determine the number of topics [17].

To verify the data pre-processing, a simple would cloud was created to see a visual representation of the most frequently occurring words in the processed text, offering insights into the thematic emphasis and recurring topics within the dataset. This also helped in determining whether further pre-processing is needed. Generally, the word cloud enhances the exploratory analysis of textual data, providing a visually intuitive representation of key themes and topics. A thorough understanding of the data is essential to identify any necessary pre-processing steps before training the model to ensure it performs optimally.

Figure 2: Generation of processed text descriptions

*D. **Modelling (LDA)***

The project selected Latent Dirichlet Allocation (LDA) as the primary topic modelling technique. Specifically, LDA was selected as the primary topic modeling technique, informed by studies such as Blei et al.[18] and Griffiths & Steyvers [19]. LDA, a generative probabilistic model, represents documents as mixtures of topics, each characterized by a distribution over words.

Key factors influencing the choice of LDA include its interpretability, flexibility, and scalability. According to Blei et al.[18], LDA offers interpretable topics, aiding in the extraction of meaningful insights from text collection. This interpretability aligns with the project's objectives, where understanding topic distributions is most important for decision-making.

Additionally, LDA exhibits computational efficiency and scalability to large datasets, as highlighted by recent studies (e.g., Hoffman et al., 2010)[20]. This scalability ensures that the modeling approach remains feasible as the dataset size or complexity increases.

While alternative topic modeling methods such as Vector Space Model (VSM), Latent Semantic Analysis/Indexing (LSA/LSI), and Probabilistic Latent Semantic Analysis (PLSA) are also prevalent, recent research (e.g., Boyd-Graber et al., 2009)[21] underscores LDA's superiority in various contexts due to its interpretability and flexibility.

IV. EVALUATION

In the evaluation of the LDA model, it's essential to consider the inherent variability in its results, stemming from the stochastic nature of the algorithm and the random initialization of topic assignments. Despite identifying 23 as the optimal number of topics through iterative processes, the LDA model consistently generates slightly different outputs with each run due to these factors. To ensure reproducibility and consistency across runs, a solution was implemented by setting a seed for the random number generator used in the LDA algorithm. This deterministic initialization ensures that the randomization process remains consistent, leading to reproducible results across multiple executions.

By addressing this variability, confidence can be placed in the analysis and interpretation of the model's performance, knowing that the observed differences are not simply artifacts of randomization but rather meaningful variations in the underlying structure of the data.

To gain insights into the underlying themes present in the dataset, LDA was employed to uncover distinctive topics and visualize them in a comprehensible manner. The bar chart visualization below in *Appendix* represents the identified topics derived from the collection of documents. Each bar chart corresponds to a specific topic, with the horizontal axis depicting the probability of each word belonging to that topic. The length of each bar reflects the probability of the associated word being part of the topic. Words with higher probabilities are more representative of the topic.

As well as a visualization tool called pyLDAvis was utilized, and it enables interactive visualization of topic models, providing a comprehensive view of the topics, their interrelationships, and the distribution of terms within each topic. As shown below in Figure 2, the resulting visualization offers an interactive display of topics, allowing users to explore the composition of each topic, the prevalence of terms, and the similarity between topics. Each bubble within the visualization represents a topic, and the size of the bubble corresponds to the significance or prevalence of the topic within the corpus of documents. Larger bubbles indicate topics that are more significant or prevalent in the collection. This significance is determined by factors such as the frequency of topic keywords and their distribution across documents.

In order to delve into the topics extracted by the Latent Dirichlet Allocation (LDA) model, an exploration of the topics and their associated keywords was conducted. This process aids in understanding the underlying themes present within the corpus of documents. Each topic, identified by a unique topic ID, was examined in detail. For each topic, the associated keywords, indicative of the predominant themes encapsulated by the topic, were displayed. The dominance of topics within the corpus was assessed based on their size within the pyLDAvis visualization. This provided insights into the significance of each topic in relation to the overall dataset.

To facilitate the organization and categorization of companies based on their thematic relevance, an association between companies and dominant topics identified by the LDA algorithm was established. This process enables the identification of related companies within topic related contexts. The output provides insights into the distribution of companies across dominant topics. Each topic is associated with a set of related companies, indicating thematic similarities among these entities.

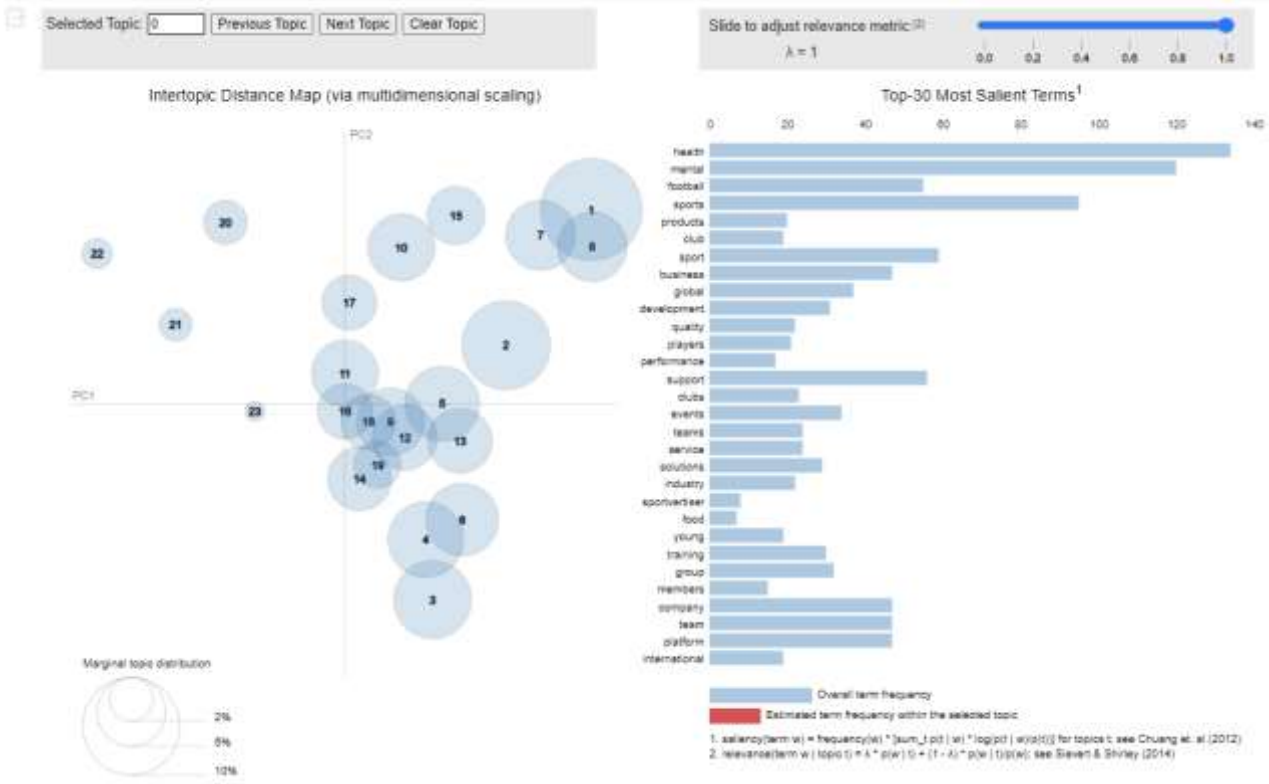


Figure 3: Topic Visualization

Topic ID	Company ID
0	{128, 35, 36, 42, 11, 13, 51, 85, 118, 158, 94}
1	{34, 68, 71, 167, 144, 83, 91, 188, 125}
2	{72, 187, 101, 102}
3	{113, 193, 165, 46}
4	{191, 199, 76, 16, 179, 181, 56, 60, 152, 127}
5	{39, 135, 105, 149, 89, 25, 63}
6	{65, 161, 163, 108, 141, 18, 52, 21, 84, 87, 117, 151, 154, 157}
7	{32, 134, 8, 40, 137, 45, 110, 156, 115, 186, 124, 189, 95}
8	{138, 109, 79}
9	{104, 114}
10	{17, 129, 86, 111}
11	{69, 107, 140, 173, 142, 80, 176, 177, 20}
12	{2, 195, 38, 73, 15, 112, 81, 148, 180, 123, 92}
13	{0, 64, 194, 106, 120, 175, 145, 53, 22, 119, 24, 155, 28, 62, 31}
14	{97, 67, 133, 6, 139, 44, 77, 90, 61}
15	{162, 185, 9, 12, 143, 121, 58, 190}
16	{66, 99, 164, 10, 170, 48, 49, 182, 184, 122, 59, 126}
17	{1, 4, 132, 43, 82, 55, 23, 26, 27}
18	{96, 192, 3, 131, 5, 7, 103, 50, 116, 54, 57}
19	{100, 37, 198, 136, 172, 19, 183, 93}
20	{98, 197, 41, 171, 146, 147, 178, 30, 159}
21	{160, 33, 196, 70, 74, 75, 150}
22	{130, 166, 168, 169, 78, 14, 47, 174, 88, 153, 29}

Table 1. Related companies for each topic ID

In the process of identifying the most optimal company based on topic relevance scores, it's essential to acknowledge the potential variability in results observed across different runs of the algorithm. The relevance scores are computed based on the distribution of topics across companies within the dataset. Many machine learning algorithms, including Latent Dirichlet Allocation (LDA), involve stochastic elements during initialization or training.

In the analysis, the most optimal company was determined to be Company ID 22 and Topic 13, with a total relevance score of 0.997. This score indicates how strongly the company is associated with the dominant topic. This company demonstrated the highest overall relevance to the topics extracted by the LDA model in the dataset.

And Topic 13 contains those keywords such as health, mental, support, young, football, care, sport, community, group, work. After conducting a thorough analysis, it has been determined that Athletes for Hope is the most optimal organization. According to their official website www.athletesforhope.org, Athletes for Hope(AFH) is an organization dedicated to empowers athletes to engage with charitable causes, increase public awareness of their efforts, and inspire others to give back. AFH vision is a world where athletes recognize their potential to make a positive impact and actively contribute to causes they care about, breaking down barriers and inspiring positive change. This aligns closely with Inter7's commitment to using sports as a platform for social good, promoting community engagement, charitable giving, and positive social impact.

To identify and evaluate the top recommended companies within the dataset based on Company ID 22's keyword occurrences and their associated topics obtained from LDA model. Top recommended companies along with the count of keyword occurrences for each company and the companies are ranked based on the frequency of keyword occurrences within their respective documents. Additionally, for each recommended company has associated topics and their probabilities from the LDA model, probabilities exceeding 0.9 for each company, focusing on high-confidence associations. Among the top recommended companies, Company ID 37 is highlighted with 9 keyword occurrences. It is associated with Topic 19 with a probability exceeding 0.9, indicating a strong thematic alignment. Similarly, Company ID 20, 28, and 39 and demonstrates each 8 keyword occurrences and is linked to Topic 11, 13, and 5, with high probability, suggesting a significant thematic focus. As well as, Company ID 23, 141, and 175 are highlighted with 7 keyword occurrences respectively to Topic 17, 6, and 13.

This validation results provide empirical evidence of the thematic alignment between identified optimal company topic IDs and top recommended companies. This process enhances confidence in the accuracy and reliability of the analysis outcomes, enabling stakeholders to make informed decisions and leverage thematic insights for strategic initiatives.

As shown below, bar chart visualizing keyword occurrences for the top recommended companies.

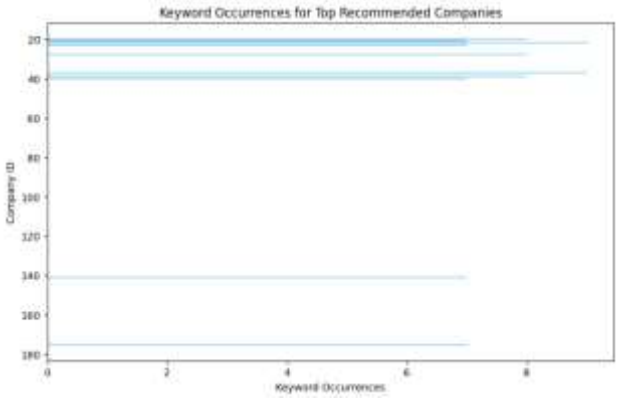


Figure 4: Keyword occurrences for the recommended companies

V. CONTRIBUTIONS AND FINDINGS

The study introduces a novel application of topic modelling in marketing strategy development, specifically focusing on identifying potential business partnerships in the charity, mental health, and sports sectors. By leveraging

Latent Dirichlet Allocation (LDA), the project explored hidden patterns within textual data from 200 companies, providing insights into thematic alignment and partnership opportunities.

The project contributes to methodological advancements in topic modelling by incorporating automated techniques for determining the optimal number of topics and evaluating the effectiveness of different modelling approaches. By addressing challenges such as data preprocessing and model interpretation, the study enhances the robustness and reliability of the analysis outcomes.

Through the identification of dominant topics and associated companies, the study offers actionable insights for strategic decision-making in marketing development. By recommending potential business partnerships based on thematic alignment and topic probabilities, the project facilitates informed collaboration and market positioning strategies for Inter7's.

The validation of thematic alignment between optimal company topic IDs and top recommended companies provides empirical evidence of the effectiveness of the analysis methodology. By demonstrating consistency between identified topics and real-world company profiles, the study enhances confidence in the accuracy and reliability of the analysis outcomes.

Generally, from the number of topics selected for modelling, which is 23 determined by the coherence score technique. Topic 13 was the dominant topic related to the most optimal company (ID 22) with the a probability of 0.6029. Among other companies, this one is most recommended to help Inter7s refine their marketing efforts in support of their league expansion. With the bubble of the topics, it if found that health, mental, support and care suggest a connection to mental health. The topics young, community, and group provided insights that companies in this bracket focus on a younger demographic and have a community aspect related to health and support. Additionally, football and sport keywords provided insights that in expanding leagues, sports psychology could be considered.

The project provides inter7s a data-driven approach to explore the intersection of mental health and sports as well as an exhaustive list of recommended companies for marketing stratification. Based on the results, the topics showed that there is potential intervention that can be done related to education and companies that deal with people with special needs (disabilities).

VI. Conclusions And Futurework

In conclusion, the study underscores the potential of topic modelling, particularly Latent Dirichlet Allocation (LDA), as a valuable tool for marketing strategy development and business partnership identification. By applying LDA to analyze textual data from diverse companies, the project successfully identifies prominent topics, extracts relevant keywords, and recommends potential partnerships based on thematic alignment. The findings validate the efficacy of the approach and highlight the importance of leveraging advanced data analytics techniques in strategic decision-making processes.

The literature showed that Interval Semi-supervised LDA (ISLDA) is superior for topic extraction in qualitative studies, Mainly because it is suitable for predefined set of keywords that are mapped to specific intervals of topic assignments, allowing for an extraction of topics related to the keywords. Therefore this model can be considered in the future given a large dataset is provided as that's the ideal data for the model to perform better than LDA. To measure how semantically coherent a model is, or how well the semantics of the corpus have been captured, perplexity could be considered, although the literature shows that this might be a bad measure. Additionally, for additional validation topic model diagnostics could be performed to assess the domain relevance via topic alignment, as validating topic models is a challenge that solely relies on human interpretability.

References

[1] M. Chebil, R. Jallouli, and M. A. B. Tobji, "Clustering Social Media Data for Marketing Strategies: Literature Review Using Topic Modelling Techniques," (in eng), Journal of Telecommunications & the Digital Economy, vol. 12, no. 1, pp. 510-537, 2024, doi: 10.18080/jtde.v12n1.889.

[2] L. E. George and L. Birla, "A Study of Topic Modeling Methods," in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 14-15 June 2018 2018, pp. 109-113, doi: 10.1109/ICCONS.2018.8663152.

[3] O. Ishmael, E. Kiely, C. Quigley, and D. McGinty, "Topic Modelling using Latent Dirichlet Allocation (LDA) and Analysis of Students Sentiments," in 2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE), 28 June-1 July 2023 2023, pp. 1-6, doi: 10.1109/JCSSE58229.2023.10201965.

[4] K. Mutiara Auliya and N. Wahyu, "Enhancing Indonesian customer complaint analysis: LDA topic modelling with BERT embeddings," in Jurnal Ilmiah SINERGI vol. 28, ed: Universitas Mercu Buana, 2023, pp. 152-162.

[5] K. K. Pandey, M. Thorat, A. Joshi, D. S, A. Hussein, and M. B. Alazzam, "Natural Language Processing for Sentiment Analysis in Social Media Marketing," in 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 12-13 May 2023 2023, pp. 326-330, doi: 10.1109/ICACITE57410.2023.10182590.

[6] B. Song and T. J. Kim, "Identifying Recent Research Topics in Post-Quantum Cryptography via Topic Modelling," in 2023 14th International Conference on Information and Communication Technology Convergence (ICTC), 11-13 Oct. 2023 2023, pp. 181-184, doi: 10.1109/ICTC58733.2023.10392726.

[7] Y. Stepaniak, V. Vysotska, O. Markiv, L. Chyrun, S. Chyrun, and L. Pohreliuk, "Technology of Text Content Topic Classification Based on Machine Learning Methods," in 2023 IEEE 5th International Conference on Advanced Information and Communication Technologies (AICT), 21-25 Nov. 2023 2023, pp. 121-126, doi: 10.1109/AICT61584.2023.10452704.

[8] C.-F. Ni, R. Lundblad, C. Dykeman, R. Bolante, and W. Łabuński, "Content analysis of psychological first aid

training manuals via topic modelling," (in eng), European Journal of Psychotraumatology, vol. 14, no. 2, pp. 1-11, 2023, doi: 10.1080/20008066.2023.2230110.

[9] J. Lande, A. Pillay, and R. Chandra, "Deep learning for COVID-19 topic modelling via Twitter: Alpha, Delta and Omicron," (in eng), PLoS ONE, vol. 18, no. 8, pp. 1-26, 2023, doi: 10.1371/journal.pone.0288681.

[10] V. Taecharungroj and I. S. Stoica, "Assessing place experiences in Luton and Darlington on Twitter with topic modelling and AI-generated lexicons," in Journal of Place Management and Development vol. 17, ed: Emerald Publishing Limited, 2024, pp. 49-73.

[11] R. Teh Faradilla Abdul, S. Raudzatul Fathiyah Mohd, B. Alya Geogiana, and N. Norshita Mat, "Topic modelling analysis of depression text message therapy: A preliminary study," in Journal of Computing Research and Innovation vol. 9, ed: Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perlis, 2024.

[12] R. L. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. R. Namburu, Data Mining for Scientific and Engineering Applications, electronic resource.

[13] Pinsent Masons, "Scraping the surface: the legality of screen scraping in Ireland," Retrieved from Lexology website, Mar. 11, 2024.

[14] Verma, Pragya. "Data Cleaning for Textual Data: A Journey from Madness." Analytics Vidhya, Mar. 30, 2022.

[15] E. P. Giachin, "Phrase bigrams for continuous speech recognition," in 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 1995, pp. 225-228 vol.1, doi: 10.1109/ICASSP.1995.479405.

[16] T. Nian and A. Sundararajan, "Social media marketing, quality signaling, and the goldilocks principle," Information Systems Research, vol. 33, no. 2, pp. 540-556, 2022.

[17] A. Thielmann, C. Weisser, T. Kneib, and B. Säfken, "Coherence based Document Clustering," in 2023 IEEE 17th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2023, pp. 9-16, doi: 10.1109/ICSC56153.2023.00009.

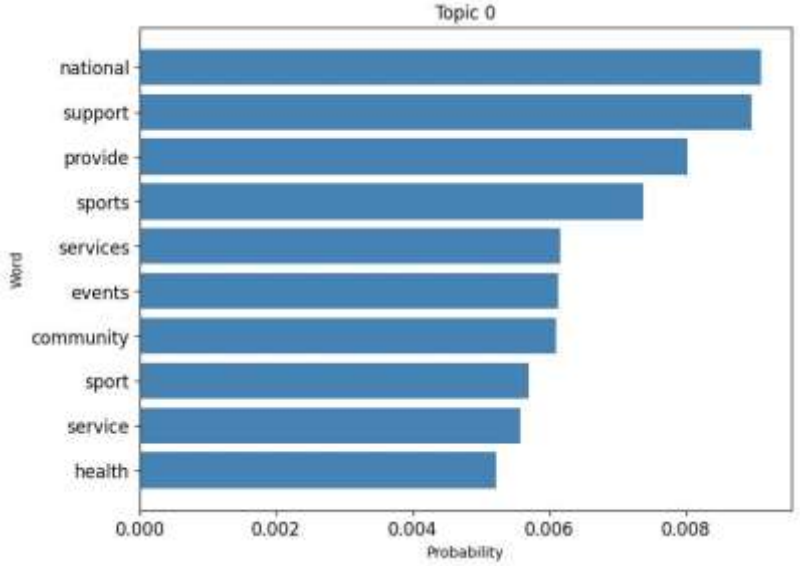
[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

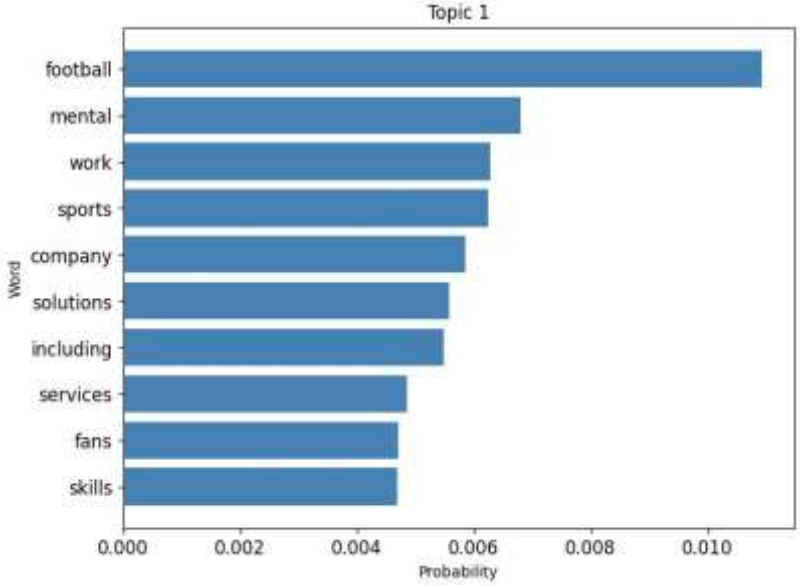
[19] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences, vol. 101, no. 1, pp. 5228-5235, 2004.

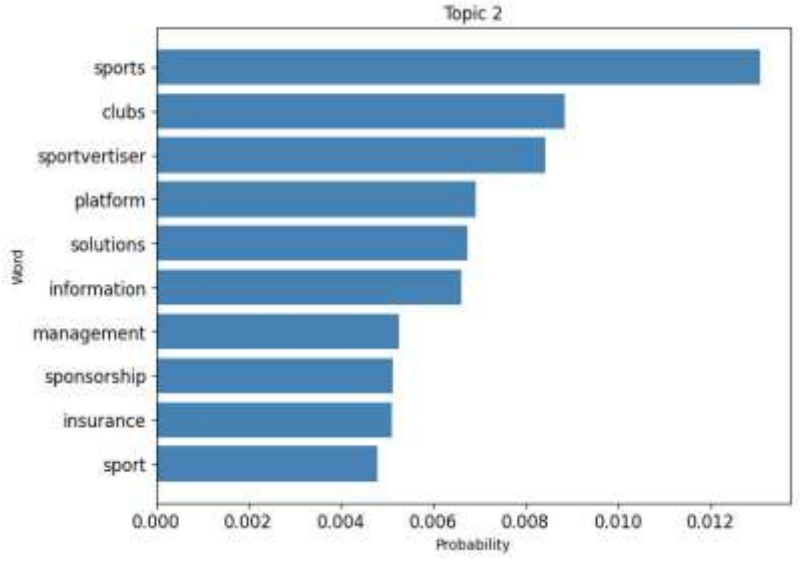
[20] M. Hoffman, D. M. Blei, and F. R. Bach, "Online learning for latent Dirichlet allocation," Neural Information Processing Systems, 2010.

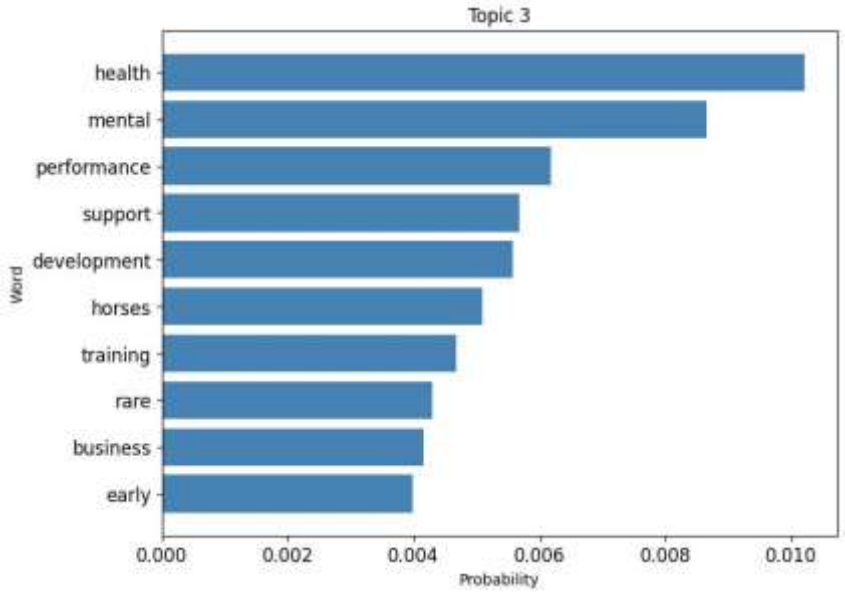
[21] J. L. Boyd-Graber and D. M. Blei, "Multilingual topic models for unaligned text," International Conference on Machine Learning, 2009.

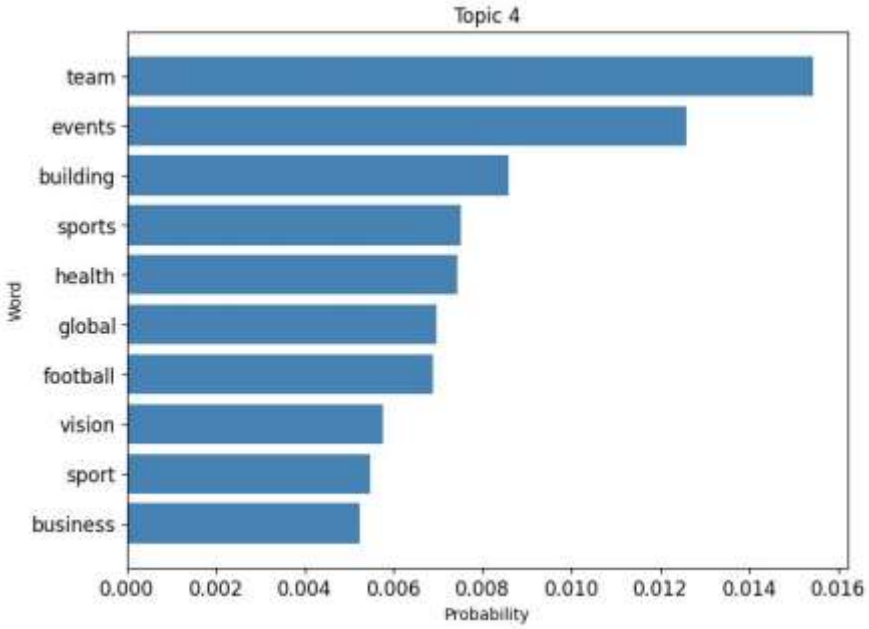
Appendix

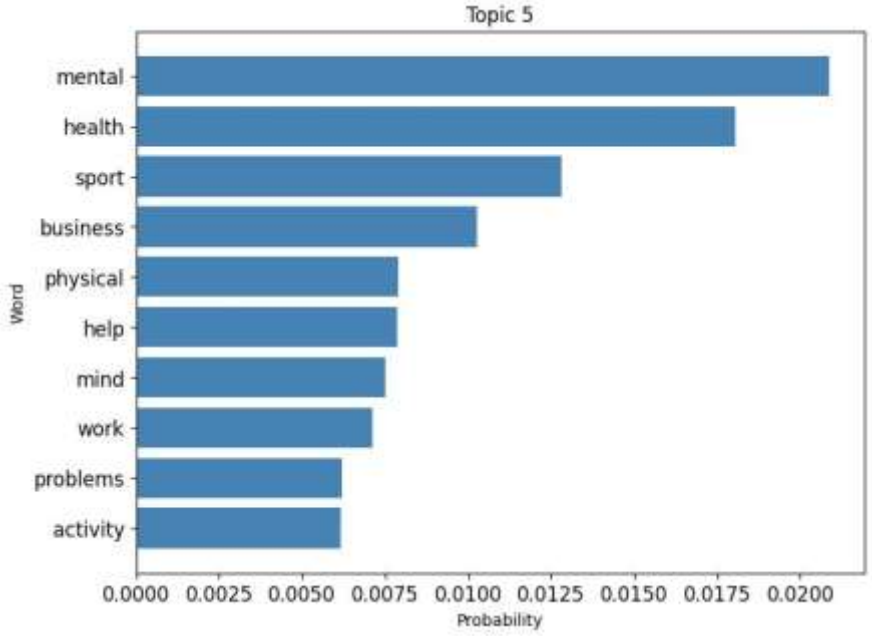

Topic 0
national
support
provide
sports
services
events
community
sport
service
health
Word
0.000
0.002
0.004
0.006
0.008
Probability

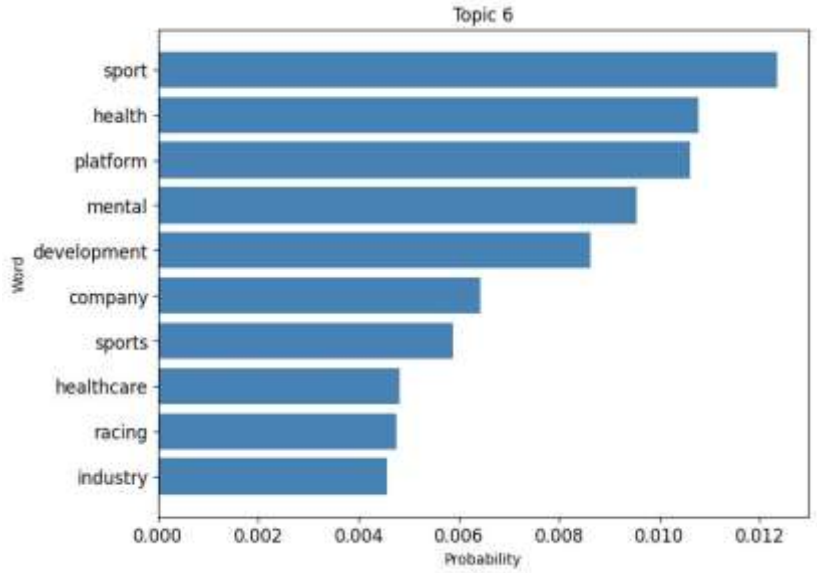

Topic 1
football
mental
work
sports
company
solutions
including
services
fans
skills
Word
0.000
0.002
0.004
0.006
0.008
0.010
Probability

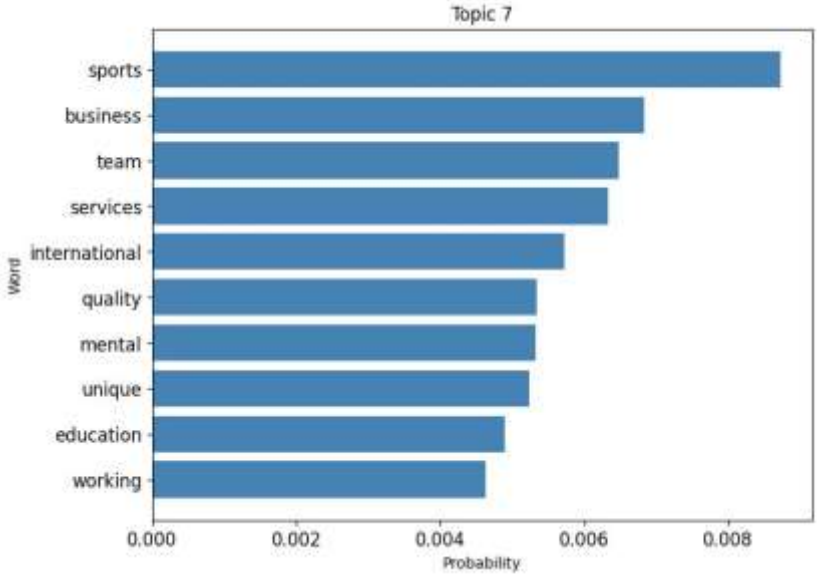

Topic 2
sports
clubs
sportvertiser
platform
solutions
information
management
sponsorship
insurance
sport
Word
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Probability

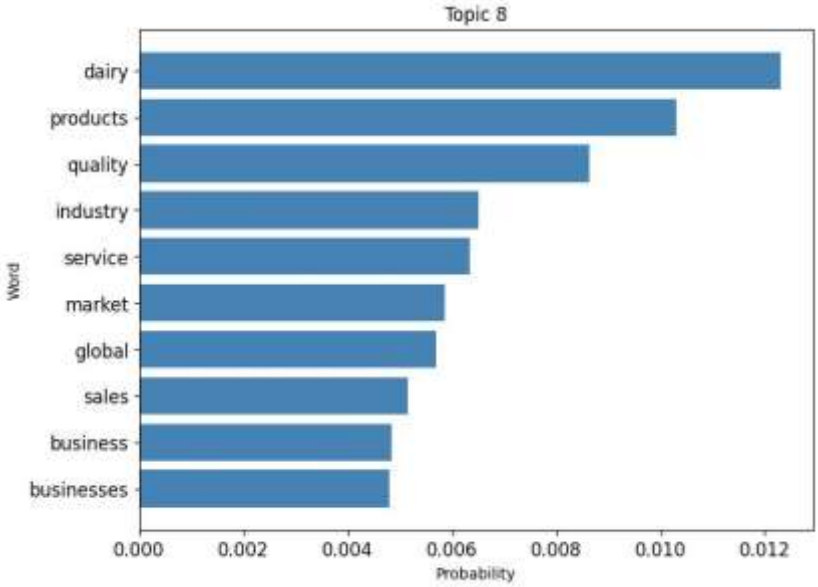

Topic 3
health
mental
performance
support
development
horses
training
rare
business
early
Word
0.000
0.002
0.004
0.006
0.008
0.010
Probability

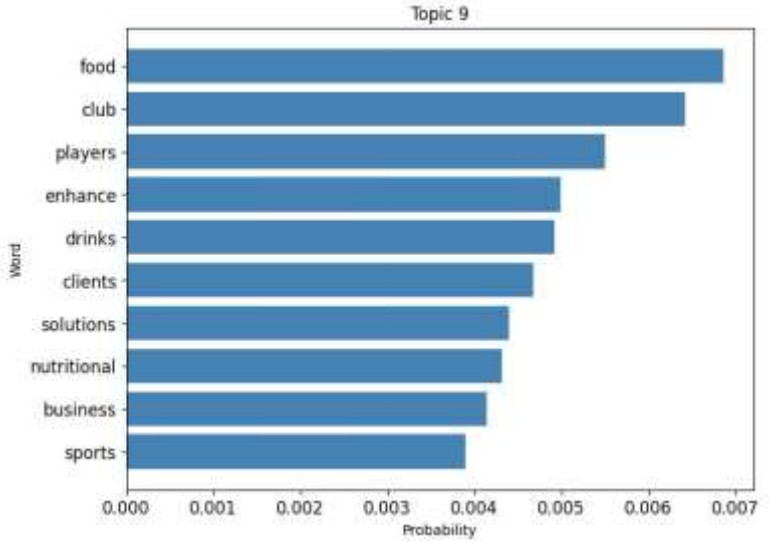

Topic 4
team
events
building
sports
health
global
football
vision
sport
business
Word
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
Probability

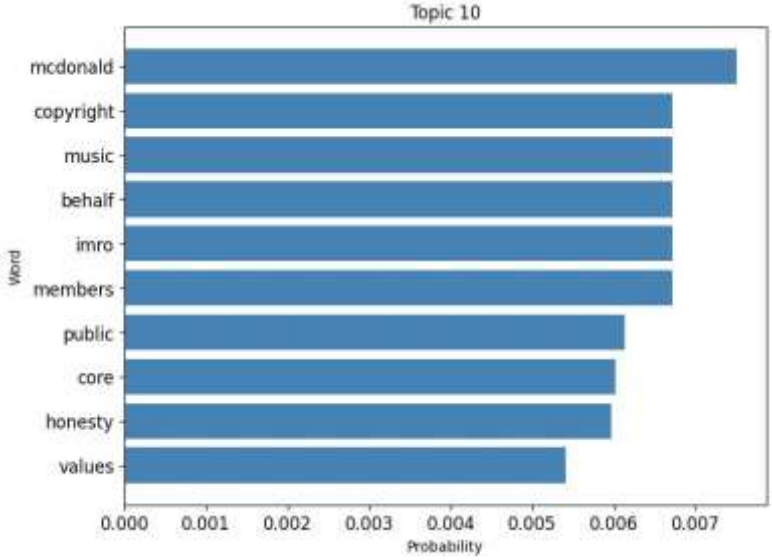

Topic 5
mental
health
sport
business
physical
help
mind
work
problems
activity
Word
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0.0200
Probability

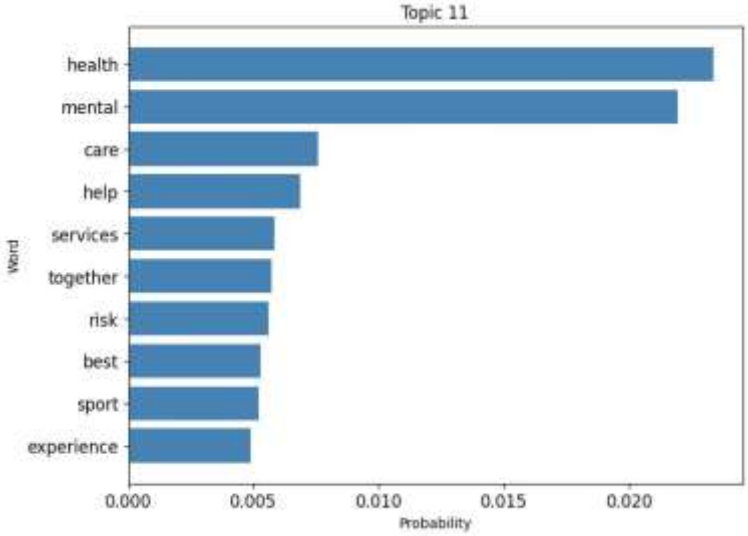

Topic 6
sport
health
platform
mental
development
company
sports
healthcare
racing
industry
Word
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Probability

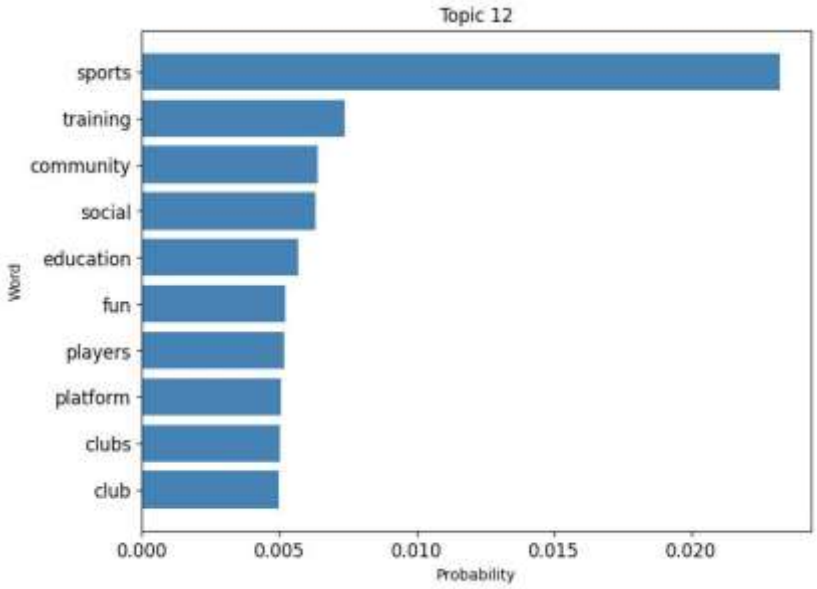

Topic 7
sports
business
team
services
international
quality
mental
unique
education
working
Word
0.000
0.002
0.004
0.006
0.008
Probability

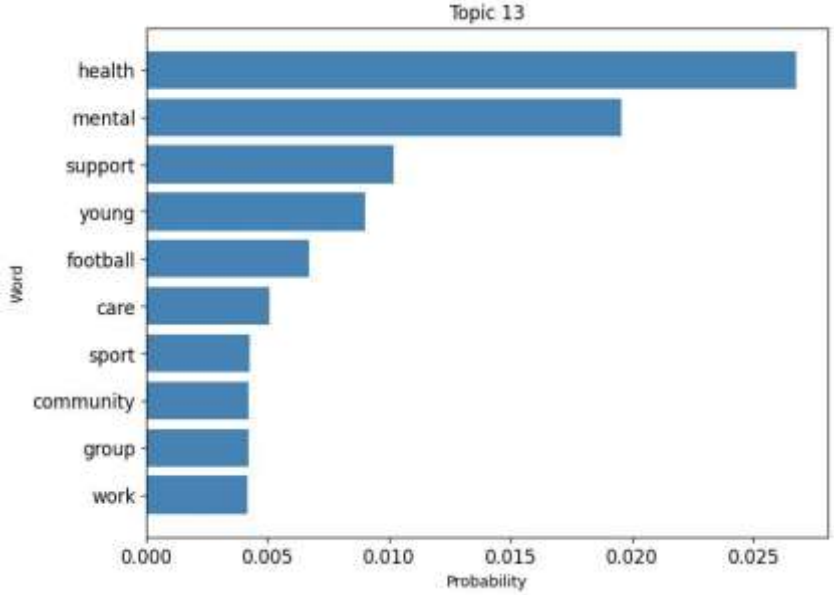

Topic 8
dairy
products
quality
industry
service
market
global
sales
business
businesses
Word
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Probability

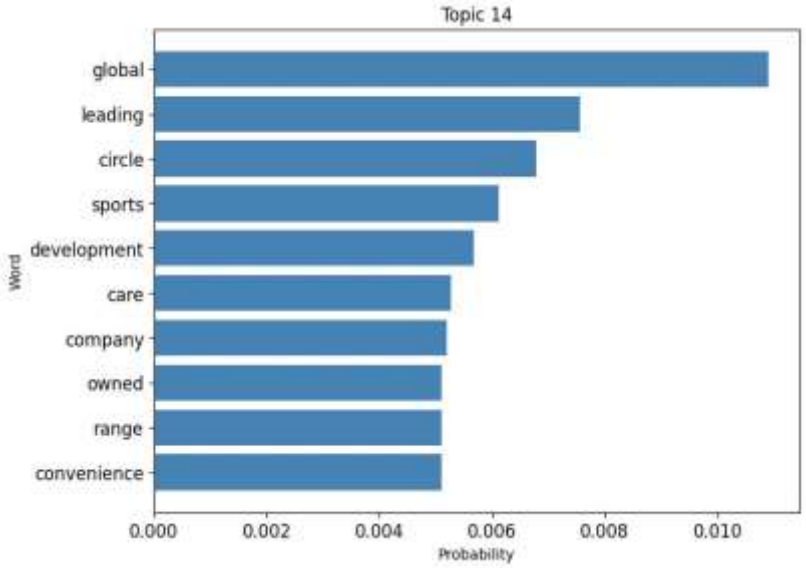

Topic 9
food
club
players
enhance
drinks
clients
solutions
nutritional
business
sports
Word
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
Probability

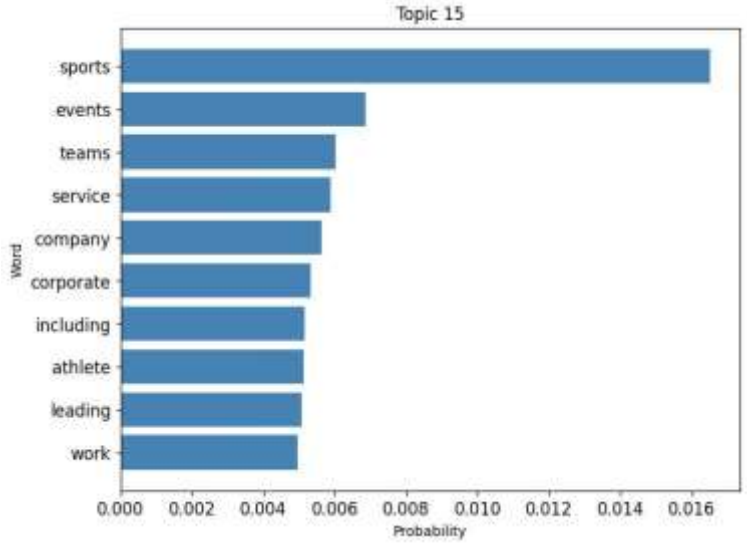

Topic 10
mcdonald
copyright
music
behalf
imro
members
public
core
honesty
values
Word
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
Probability

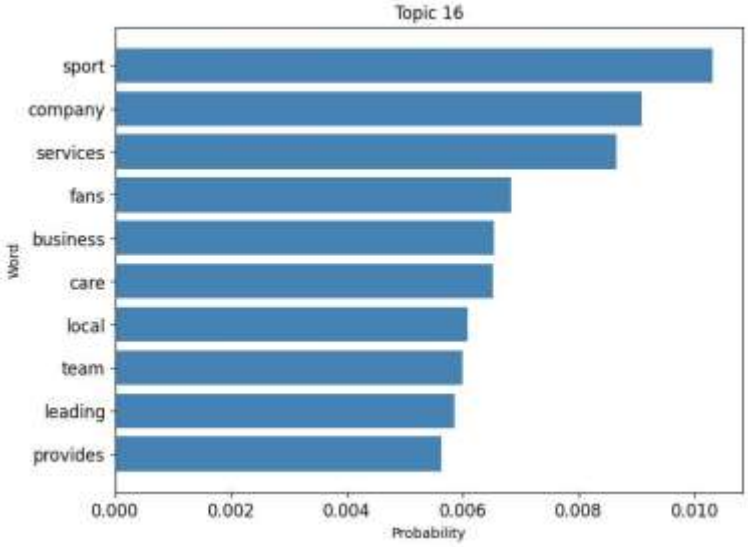

Topic 11
health
mental
care
help
services
together
risk
best
sport
experience
Word
0.000
0.005
0.010
0.015
0.020
Probability

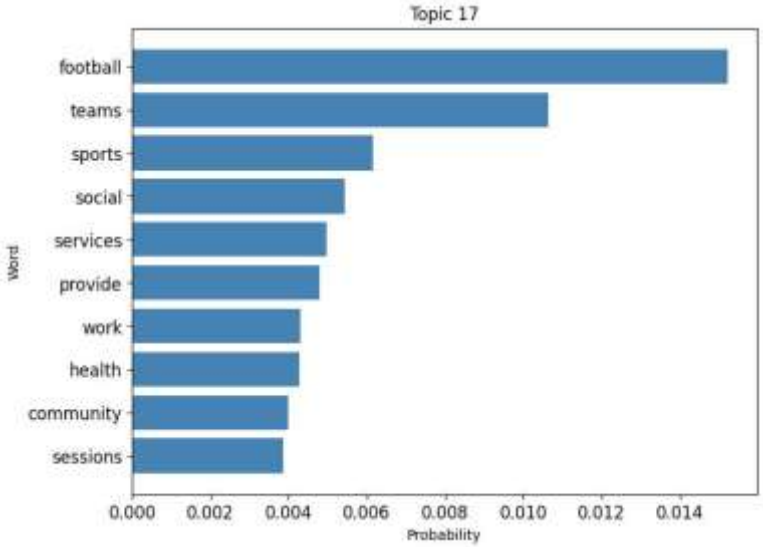

Topic 12
sports
training
community
social
education
fun
players
platform
clubs
club
Word
0.000
0.005
0.010
0.015
0.020
Probability

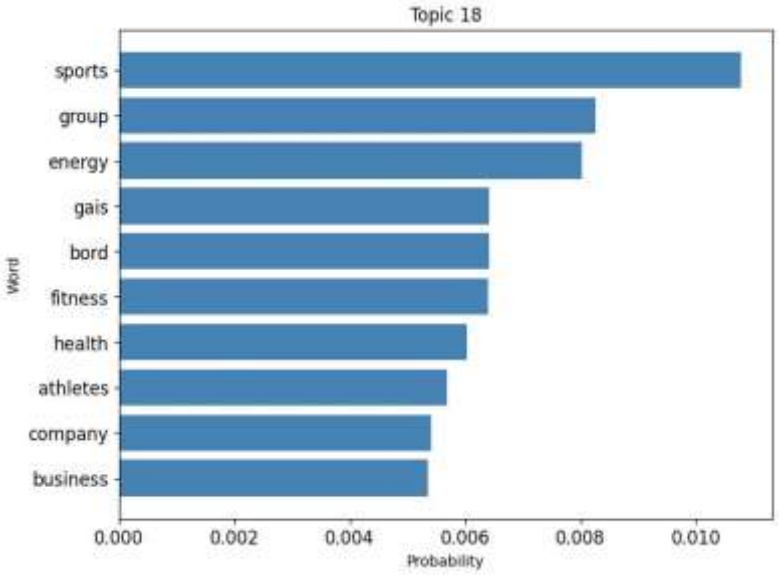

Topic 13
health
mental
support
young
football
care
sport
community
group
work
Word
0.000
0.005
0.010
0.015
0.020
0.025
Probability

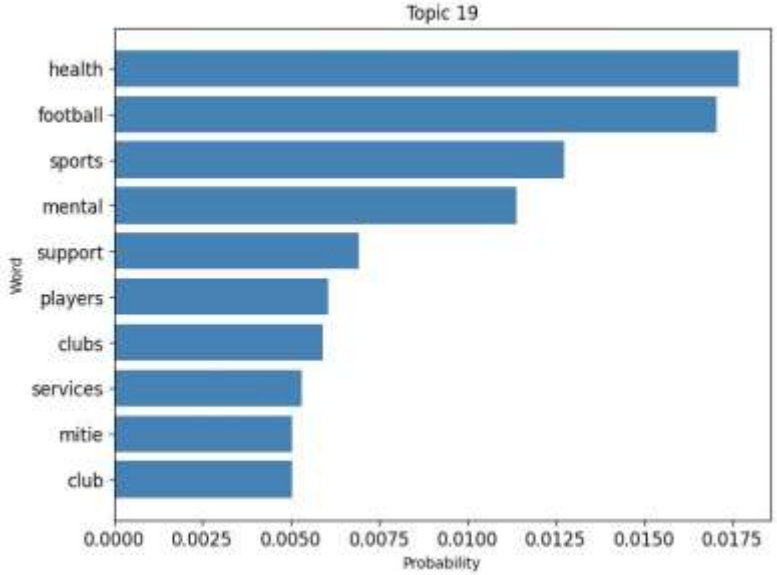

Topic 14
global
leading
circle
sports
development
care
company
owned
range
convenience
Word
0.000
0.002
0.004
0.006
0.008
0.010
Probability

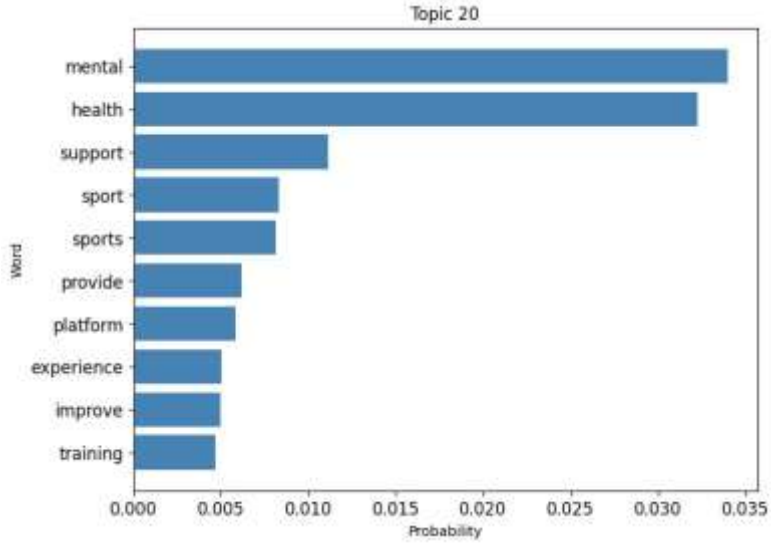

Topic 15
sports
events
teams
service
company
corporate
including
athlete
leading
work
Word
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
Probability

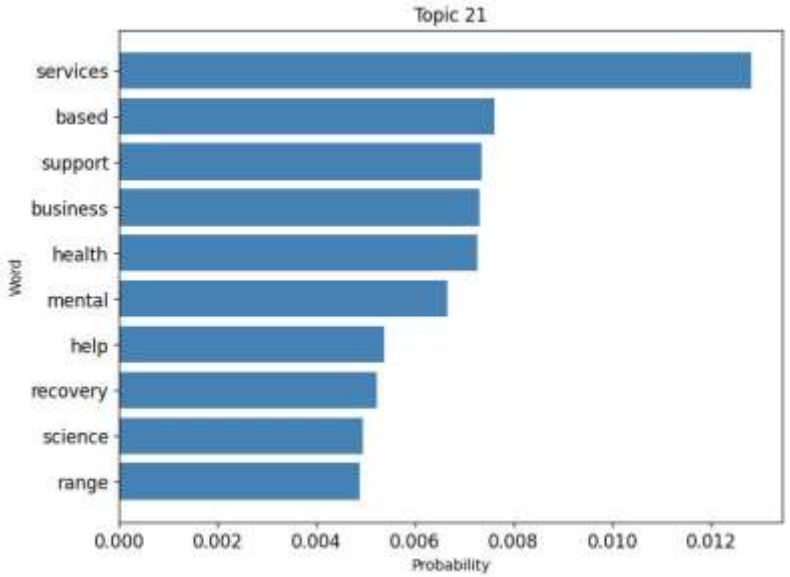

Topic 16
sport
company
services
fans
business
care
local
team
leading
provides
Word
0.000
0.002
0.004
0.006
0.008
0.010
Probability

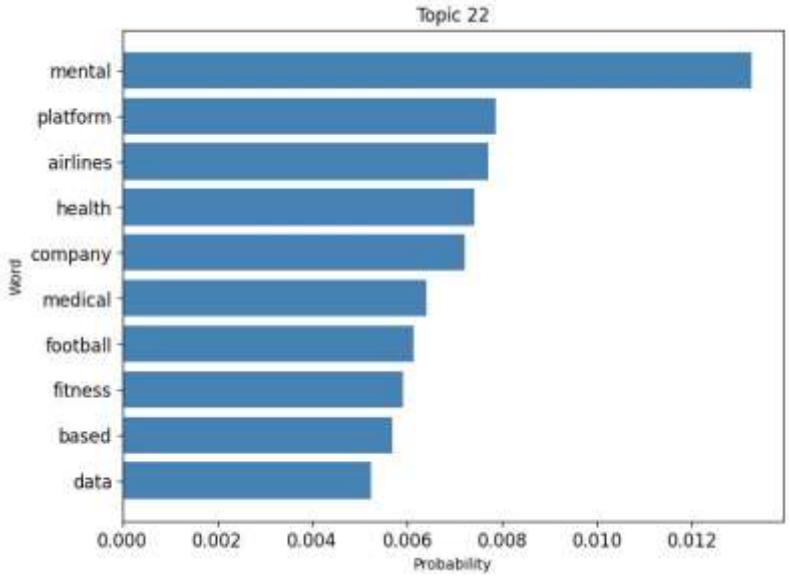

Topic 17
football
teams
sports
social
services
provide
work
health
community
sessions
Word
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
Probability


Topic 18
sports
group
energy
gais
bord
fitness
health
athletes
company
business
Word
0.000
0.002
0.004
0.006
0.008
0.010
Probability


Topic 19
health
football
sports
mental
support
players
clubs
services
mitie
club
Word
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
Probability


Topic 20
mental
health
support
sport
sports
provide
platform
experience
improve
training
Word
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Probability


Topic 21
services
based
support
business
health
mental
help
recovery
science
range
Word
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Probability


Topic 22
mental
platform
airlines
health
company
medical
football
fitness
based
data
Word
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Probability