# Emotion Detection for Image-Based Classification

Lunaba, Christian Lee
*College of Computing and Information Technology*
Manila, Philippines
lunabac@students.national-u.edu.ph

Montaño, Chleo Nicole C..
*College of Computing and Information Technology*
Manila, Philippines
montanocc@students.national-u.edu.ph

Paat, Margarete A..
*College of Computing and Information Technology*
*Manila, Philippines*
*paatma@students.national-u.edu.ph*

***Abstract*—Abstract: Understanding the job market is crucial for both employers and job seekers. This study explores job postings using topic modeling and clustering techniques to identify trends and relationships between different industries, specifically utilizing LDA, t-SNE, and UMAP. The results were then evaluated through word sampling and compared across different clustering algorithms, including K-Means, Agglomerative Clustering, Gaussian Mixture Models (GMM), DBSCAN, and Spectral Clustering, alongside Word2Vec and UMAP.**

**The researchers found that LDA combined with t-SNE and UMAP provided deeper insights into job market trends. Future work can improve upon this study by implementing rigorous evaluation metrics for validating the insights, using more advanced word embeddings such as BERTopic or Top2Vec, replacing Word2Vec with BERT, and exploring alternative dimensionality reduction techniques beyond UMAP.**

**Keywords: Job Market, Topic Modeling, Clustering, LDA, t-SNE, UMAP, BERT, Industry Trends**

## I. Introduction

The increasing demand for unique and specialized skills, along with the wide range and vast number of job postings across various platforms, makes it challenging to identify specific qualities necessary for specific roles. Understanding the job trends and required abilities is crucial for job seekers and employers alike. According to recent employment descriptions, skill-based hiring is becoming a significant factor in recruitment, highlighting the importance of data-driven insights in career decision-making.

The primary challenge addressed in this research is the complexity of navigating job postings and identifying essential skills for specific positions, which jobs are most relevant and which jobs are connected to each other. Especially in this day and age where job vacancies are all over the internet. Our proposed solution involves collecting and analyzing LinkedIn job postings using web scraping techniques, extracting key terms, and applying clustering algorithms to group words with similar contexts. This will provide a structured interpretation of job trends across different industries.

The significance of this research lies in its ability to assist job seekers in aligning their skill sets with industry demands, improving their chances of employability with the information regarding their job interest with the use of clustering results. Additionally, this study benefits employers by helping them refine job descriptions and understand labour market trends, while the job seekers have the privilege to hone their skills in alignment. The primary users of this solution include job seekers, employers, researchers, students, and aspiring technology professionals who wish to gain insights into evolving job market trends. The findings can be applied in career counselling, planning, and job recommendation systems to bridge the gap between job seekers and potential employers.

## II. Review of Related Literature

In the late 1900s and early 2000s, individuals primarily relied on newspapers and online job listings to find employment. Early online job boards like Monster.com used keyword-based search mechanisms that lacked contextual understanding, leading to irrelevant search results [1]. A major breakthrough occurred in 2013 with the development of Word2Vec, a neural network-based model introduced by Mikolov et al., which represented words as vectors to improve semantic similarity detection and enhance word relationship understanding [2]. Following this, platforms like LinkedIn and Indeed integrated machine learning algorithms to recommend jobs based on user behavior, interests, and profile data, significantly improving job search relevance [3], [4].

LinkedIn introduced a market-aware skill extraction system that enhances job recommendations and skill suggestions for job seekers and employers. This system leverages natural language processing (NLP) techniques to recognize variations of similar skills, rank their importance, and improve the performance of job-matching algorithms [5].

Word2Vec, a widely used word embedding technique, applies two iteration-based methods: Continuous Bag of Words (CBOW) and Skip-Gram. CBOW predicts a target word from its surrounding context, while Skip-Gram predicts surrounding words given a central word [2]. Studies have demonstrated that Word2Vec effectively captures contextual relationships between words, making it useful for various NLP applications, including job recommendations [6].

The study ”Approach to the Use of Language Models BERT and Word2Vec in Sentiment Analysis of Social Network Texts” compared Word2Vec and BERT in text vectorization. Word2Vec creates a dictionary from a dataset, learns word co-occurrence patterns, and adjusts vectors accordingly, whereas BERT processes entire text segments to capture context at a deeper level [7]. Another study, ”Evaluating Word Embedding Models: Methods and Experimental Results,” awarded the Sadaoki Furui Prize Paper Award, analyzed different word embedding models. It highlighted the Continuous Bag-of-Words (CBOW) and Skip-gram models used in Word2Vec, where CBOW predicts a missing word based on surrounding words, while Skip-gram predicts surrounding words based on a given word. The study also compared Word2Vec with GloVe, which focuses on word co-occurrence, and ngram2vec, which enhances phrase understanding. The conclusion emphasized that while both models perform well in general, newer models like ngram2vec improve phrase understanding [8].

In this study, clustering is used. Clustering is a technique to group similar data points. It organizes a set of objects into groups known as clusters so that data points in a group are more similar compared to those in other groups. Spectral clustering is a variant of the clustering algorithm that uses mathematical tools to identify similarity between data points to decide how to group them. It works for even irregular shapes of clusters, meaning more complex [9].

Another paper entitled “A Comprehensive Survey of Clustering Algorithms,” identifies the uses of some known clusters. First is Hierarchical Clustering, a broad aspect where it is about how clusters are formed or creating a structure of nested clusters. The hierarchical relationship is the process of starting with each point or small clusters and then merging the closest clusters together, which is a process called Agglomerative Clustering, a type of hierarchical clustering [10]. Another one, which is K-means, is a partition-based clustering algorithm and focuses on dividing data into a fixed number of clusters by finding the center of each cluster or the average of data points to the nearest cluster center until the centers no longer change. Fourthly, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is where data points in regions of high density are grouped into clusters, while the points found in low-density regions, outliers, are not considered in the clusters [10]. It looks at the density of points using two parameters: Radius and Minimum Points. Radius is the maximum distance to consider for neighboring points, and Minimum Points is the minimum number of points required to form a cluster. Last is the Gaussian Mixture Model, generated from normal distributions. Each cluster has one of these Gaussian distributions, and data points are assigned to clusters based on their probability of coming from the different distributions or calculating the probability that a data point belongs to each cluster [10].

Additionally, since modern datasets exist in high-dimensional spaces, datasets with a large number of features or variables, dimensionality reduction is essential for managing complexity [11]. t-SNE (t-Distributed Stochastic Neighbor Embedding) is used for visualization and clustering. t-SNE is a nonlinear dimensionality reduction algorithm that groups similar words in a lower-dimensional space, helping to organize job-related terms into clusters and suggest relevant keywords [12]. This helps the use of Word2Vec in improving job-related term associations and recommendations.

Furthermore, ”Dimensionality Reduction and Classification of Hyperspectral Remote Sensing Image Feature Extraction” paper includes different dimensionality reduction and supervised machine learning algorithms to determine suitable combinations of classification and dimensionality reduction methods for images [13]. It focuses on feature extraction and includes Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). SVD is a linear method for reducing the number of features in a dataset, which is similar to PCA, an algorithm that finds the key directions or principal components that hold the most important information. SVD, on the other hand, focuses on the singular values or key components that represent the majority of the data, keeping only the essential details. According to the authors, SVD provides the best classification accuracy for hyperspectral data [13].

UMAP (Unified Manifold Approximation and Projection), a non-linear dimensionality reduction technique for mapping the original data to a low-dimensional space, is being used as well, to determine if it can remove redundant data and maintain good classification accuracy for the high-dimensional characteristics of hyperspectral data (information used in images to allow detailed analysis) [13]. This makes it more useful in data clustering because it finds not only relationships of the global structure but also small patterns. Results show that UMAP not only has high accuracy but also reduces running time [13]. The analysis indicates that using UMAP as a data dimensionality reduction algorithm achieves superior performance in the terrain classification of hyperspectral images, with a classification accuracy of 99.57% [13]. It concluded that dimensionality reduction not only reduces the amount of data but also achieves good classification accuracy.

Lastly, the “Latent Dirichlet Allocation” study explains that LDA (Latent Dirichlet Allocation), a model where different levels of data, such as document, topic, or word, interact with each other. It assumes that each document is a combination of multiple topics and that each topic has a specific set of words associated with it [14]. The model aims to find topics in a document and estimate how likely a word is to appear in a topic by looking at the collection of documents.

Word2Vec has influenced NLP applications, including job recommendation systems, by improving the contextual representation of words [2]. Clustering algorithms and probabilistic models like LDA help identify and relate words and meanings. Reduction techniques such as t-SNE and UMAP provide significant support for clustering. Overall, these techniques help related words be searched more effectively.

## III. Methodology

### A. Data Collection

The researchers used a collection of dataset gathered from 2023-2024 using a web scraper in selenium. The data consisted of 124,000+ job postings. Each job posting includes numerous valuable attributes for both the job and the company, such as the title, description, salary, location, application URL, and work type (e.g., remote, contract). Additionally, separate files contain details on benefits, required skills, and associated industries. The dataset contains a total of 3 folders which are companies, jobs and mappings. Companies folder contains descriptions for company industries, specialties an employee counts. The jobs folders contain job benefits, industry of the job, skills related to the job and salaries of the given job. Because the dataset only contains data from 2023-2024, the researchers used the script provided by the notebook to scrape data for 2025. An additional 2000+ Job Postings were scrapped for a duration of 2 days.

### B. Data Pre-Processing

Since the data was distributed across multiple tables, the researchers first merged them using their respective table IDs. The skills table was integrated with the job postings, with each skill separated by “”and”“ to optimize the algorithm’s performance. Next, the benefits and company industries tables were merged with the job postings. Finally, the company specialties table was also merged and formatted similarly to the skills table.

Once the dataset was fully merged, the researchers compiled the data into a single paragraph to create a structured job posting. Only the following columns were retained: "company_name", "title", "description", "max_salary", "pay_period", "location", "med_salary", "min_salary", "formatted_work_type", "remote_allowed", "application_type", "expiry", "formatted_experience_level", "skills_desc", "posting_domain", "sponsored", "work_type", "currency", "compensation_type", "normalized_salary", "zip_code", "industry_name", "skill_name", "type", "industry", and "speciality".

To ensure word embedding algorithms like Word2Vec and LDA treated salary values as a single entity, the "min_salary", "max_salary", and "med_salary" values were concatenated using an underscore (_). If a job posting was remote, the phrase "remote_allowed" was appended to the paragraph. Similarly, if a posting was sponsored, "sponsored_yes" was added. Finally, all data was concatenated to generate a structured job posting.

After parsing the job postings, WordNinja was used to separate incorrectly connected words. This ensured that words were properly treated to improve the performance of the algorithms employed by the researchers.

### C. Feature Generation, Transformation and Experimental Setup

The researchers used pandas as the main tool to parse the CSV files. They also utilized a website named Embedding Projector - Visualization of High-Dimensional Data for data visualization. This website helped visualize results for UMAP and t-SNE for both 2D and 3D. Finally, for scraping the dataset for 2025, the researchers used the following GitHub repository: GitHub - ArshKA/LinkedIn-Job-Scraper, which allows retrieving and storing a live stream of job postings.

After processing the data, the researchers proceeded with the experimentation, which consisted of four key steps. First, they analyzed the relationships between words within each job posting and clustered the dataset. The primary algorithm used for this purpose was LDA (Latent Dirichlet Allocation).

Next, to visualize the resulting embeddings, the researchers used different data visualization techniques. This included t-SNE, which excels at clustering closely related points. However, t-SNE lacks the ability to provide meaningful insights about the distance between clusters and their sizes. These pieces of information are crucial to determining the complexity of a certain industry and its relation to other industries.

To address this limitation, the researchers also used UMAP (Uniform Manifold Approximation and Projection), a powerful dimensionality reduction technique that preserves both local and global structures in the data.

The results of each data visualization technique were analyzed and compared.

For the third step, using the results from the previous step, the researchers provided valuable insights and suggestions they deemed helpful in achieving the goals of this research. These insights were based purely on the results and the researchers’ understanding of the job market, either through research or experience.

Lastly, to verify and improve upon these ideas and suggestions, the researchers used different clustering algorithms and compared them with each other using 3 metrics, chose the best clustering algorithm and compared that with the results of LDA combined with both TSNE and UMAP. Since most clustering algorithms do not have a built-in semantic analyzer like LDA, the researchers employed Word2Vec. The researchers then used UMAP again as the main data visualization technque for each clustering algorithm, mainly because of its speed and ability to preserve local similarities unlike TSNE. The overall pipeline used by the researchers for different clustering algorithms was as follows:

$$\text{Word2Vec} \rightarrow \text{UMAP} \rightarrow \text{Clustering Algorithm}$$

The clustering algorithms explored were:

- K-Means
- Agglomerative Clustering
- DBSCAN
- Spectral Clustering
- Mixture Of Gaussians

The results of each clustering algorithm were analyzed and measured based on the following evaluation metrics:

- Silhouette Score
- Davies-Bouldin Index
- Dunn Index
- Calinski–Harabasz Index

### *D. Algorithm*

The researchers used different sets of algorithms for different purposes. The main algorithm focused on by the researchers is called Latent Dirichlet Allocation (LDA). It works by analyzing word distributions across different topics. By doing so, LDA can infer semantic relationships between words. The expectation is that if certain words appear frequently across multiple documents, LDA will eventually learn that these words are connected. This allows researchers to gain insights into which skills are related to certain topics, which jobs are closely related to each other, and which requirements are associated with specific job types. These insights help in understanding the current landscape of the job market on LinkedIn. Due to time constraints, the researchers opted for this algorithm despite the availability of potentially better techniques.

**Dimensionality Reduction Techniques**

For visualization techniques, the researchers made use of t-SNE and UMAP

*1) t-SNE:* t-SNE is a nonlinear dimensionality reduction technique designed to map high-dimensional data into a lower-dimensional space while preserving local relationships between points. The algorithm works as follows:

- **Computing pairwise similarities:** It calculates probabilities that represent how similar two points are in high-dimensional space. Points that are close to each other in the higher-dimensional space should have a high probability of landing close to each other in the embedding space.
- **Early exaggeration:** At the beginning of the optimization process, t-SNE applies an early exaggeration phase, which increases the influence of local similarities. This helps create well-separated clusters in the lower-dimensional space, making the structure of the data more apparent.
- **Mapping into lower dimensions:** The algorithm optimizes a new distribution in the target space that closely resembles the original high-dimensional relationships.
- **Minimizing divergence:** It iteratively adjusts the positioning of points to minimize the Kullback-Leibler (KL) divergence between the high-dimensional and low-dimensional probability distributions, ensuring that similar points remain close while dissimilar points stay apart.

t-SNE is chaotic by nature and involves sensitive parameters such as perplexity. The researchers tested four versions of t-SNE with varying perplexities: 40, 50, 60, and 70. Given the large dataset, a high perplexity was used. However, since t-SNE primarily preserves local distances, it is effective for clustering but lacks the ability to provide meaningful insights into relationships between different clusters.

*2) UMAP:* To address this limitation, the researchers compared the results with UMAP. UMAP (Uniform Manifold Approximation and Projection) is another nonlinear dimensionality reduction technique designed to preserve both local and global structures in high-dimensional data.

UMAP works by:

- **Constructing a high-dimensional graph:** UMAP builds a weighted k-nearest neighbors (k-NN) graph, capturing both local and global similarities in the dataset.
- **Optimizing a low-dimensional representation:** The algorithm then brings this constructed map from high dimension to lower dimension, maintaining the graph as much as possible, hence the name Projection.
- **Preserving both local and global structures:** Unlike t-SNE, which primarily maintains local similarities, UMAP creates a graphical mapping between most, if not all, of the points, allowing it to preserve global similarities.

To evaluate UMAP's effectiveness, the researchers tested multiple configurations by varying the number of neighbors (*n_neighbors*) and minimum distance (*min_dist*) parameters.

**Clustering Algorithms**

The researchers compared the results of LDA with other clustering algorithms using Word2Vec and UMAP.

*3) Word2Vec:* Word2Vec is a neural network-based model that transforms words into dense vector representations, capturing semantic relationships based on word co-occurrence in a given corpus. It operates using two primary architectures:

- **Continuous Bag of Words (CBOW):** Predicts a target word based on its surrounding words, effectively capturing contextual meaning.
- **Skip-gram:** Predicts surrounding words given a target word, allowing it to model relationships even in sparse data.

By training on a large dataset of job postings, Word2Vec positioned similar words closer together in vector space. This allowed the model to recognize synonyms, industry-specific jargon, and word associations unique to job descriptions. For example, terms like ”software engineer,” ”developer,” and ”full stack” would be mapped near each other, reflecting their relatedness in the job market. Other alternatives, such as BERT, exist, but due to the computational cost, the researchers opted for Word2Vec, which offered faster processing and was suitable for multiple iterations.

*4) Clustering Methods:* Using this tool, the researchers then explored different clustering algorithms, including:

- **K-means**
- **Agglomerative Clustering**
- **DBSCAN**

- **Mixture of Gaussians**
- **Spectral Clustering**

### E. Comparison with other Clustering Algorithms

After the **LDA + UMAP** technique was done, different clustering algorithms were tested to analyze, compare, and improve upon the findings the researchers obtained. The researchers tested numerous clustering algorithms such as **K-means**, **Hierarchical clustering**, **Agglomerative clustering**, **DBSCAN**, **Mixture of Gaussians**, and **Spectral Clustering**. These algorithms were chosen with their strengths in mind to better understand where LDA performs well and where it does not.

**K-means** works well for well-separated, spherical clusters but struggles with complex shapes. Comparing LDA with K-means provides insights into whether the resulting dataset has a complex or well-separated structure. **Hierarchical clustering**, specifically **Agglomerative clustering**, is useful as it determines the best number of clusters without requiring an explicit predefined value. This helps researchers assess whether the initial assumption of ten clusters was sufficient for clustering the data.

**DBSCAN** was also used as a comparison since it is more effective for non-spherical data, unlike K-means. **Gaussian Mixture Models** were incorporated due to their ability to perform *soft clustering*, which is useful for identifying overlaps between clusters. This is particularly relevant in the job industry, where job roles and required skills often overlap. Lastly, **Spectral Clustering** was explored as it excels at handling non-convex, highly structured data, where both K-means and DBSCAN face challenges.

These clustering algorithms were paired with a word embedding algorithm known as **Word2Vec**. The results of these clustering algorithms were then evaluated using different metrics, which will be discussed in the next section.

### F. Evaluation Metrics

Most of the insights presented in this paper are highly subjective; hence, no systematic metrics were used for the first two steps of the experiments. Still, the researchers used techniques to better gauge the relevance of the results by analyzing whether they can provide valuable insights for understanding the current landscape of the job market. The main technique used was word sampling. Multiple sets of words were sampled to analyze the relationship between words in terms of (1) Industry, (2) Skills, (3) Remoteness, (4) Salary, (5) Job Title, (6) Location, and (7) Company. Various insights were then produced based on these sampled results.

For the last two steps of the experiment, multiple evaluation metrics were used to assess the clustering structure of the results without requiring ground truth labels. These types of metrics are also called **Internal Evaluation Metrics**. The metrics are as follows:

- **Silhouette Score**: Measures how similar a point is to its own cluster compared to other clusters. Ranges from -1 to 1 (higher is better).
- **Davies-Bouldin Index**: Evaluates cluster compactness and separation. Lower values indicate better clustering.
- **Dunn Index**: The ratio of the smallest inter-cluster distance to the largest intra-cluster distance. Higher values indicate better clustering.
- **Calinski-Harabasz Index**: Ratio of the sum of between-cluster dispersion to within-cluster dispersion. Higher values indicate better-defined clusters.

Obviously, these metrics are not enough on their own to determine whether a certain clustering algorithm performed better than another. More importantly, it is not enough to use these results to benchmark LDA. Hence, the researchers also analyzed whether the results produced by the clusters were insightful in nature and helpful in understanding the current landscape of the job market using the same word sampling technique. The benchmark consists of comparing the insights between LDA and the clustering techniques alongside the results of the internal evaluation metrics.

## IV. Results and Discussion

### A. LDA word samples

First, the researchers evaluated each cluster using the topics produced by LDA. Here are the top relevant words per topic:

- **Topic 0:** Technology, building, strategy, clients, performance, management, success, retail, products, development, market, media, growth, business, sales, brands.
- **Topic 1:** Full_time, market, health, financial_services, finance, products, customers, service, client, insurance, account, business.
- **Topic 2:** Service, dental, rehabilitation, provider, families, physicians, qualifications, medicine, licensed, environment, certified, therapy, physical, health_care_provider, practice, patient.
- **Topic 3:** Developing, collaborate, remote, agile, cyber, architecture, engineers, enterprise, code, perform, quality, functional, SQL, analysis, computer, science, IT, information technology.
- **Topic 4:** Software, specialist, finance, monthly, technical, material, technician, industrial, engineering, physical, mechanical, maintenance, art, design.
- **Topic 5:** Accountant, specialist, finance, laboratory, Excel, document, policies, company, billing, assistant, payroll, computer, accounts, insurance.
- **Topic 6:** Medical, government, offsite, human, sex, religion, veteran, race, color, degree, sexual, research, law, department, protected, qualified, national, federal, education, position.
- **Topic 7:** External, relationships, contract, HR, organizational, stakeholders, staff, field, office, compliance, communication, management, international.
- **Topic 8:** Products, production, market, care, merchandise, associate.
- **Topic 9:** Recruiter, health vision, attorney, litigation, competitive, law, network, law, medical, travel, dental, contract, legal, professional.

From the given topics, the researchers suggested the following industries related to each:

0) Business and Technology
1) Health and Finance
2) Medicine, Environment, and Safety
3) Data and Security
4) Engineering and Design
5) Banking
6) Legal and Relations
7) Travel or Legalities
8) Production
9) Undefined

### *B. LDA + TNSE and UMAP RESULTS*

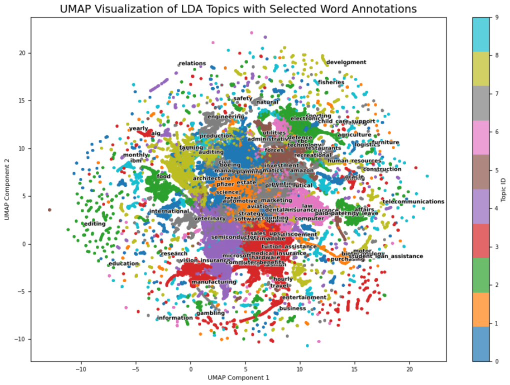

Fig. 1. LDA + UMAP

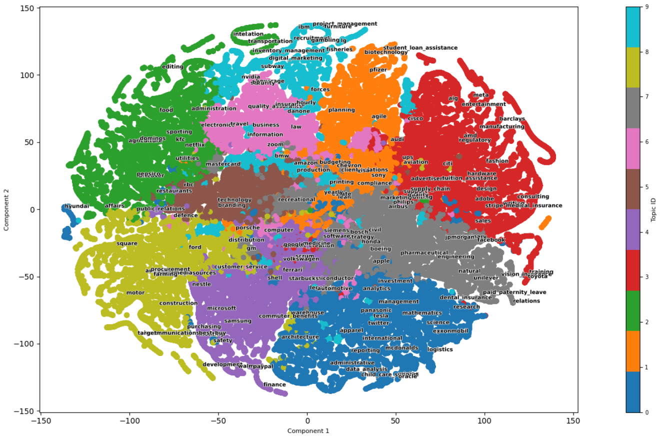

Fig. 2. LDA + TSNE

**Cluster 0: Business and Technology**

Using the results from TSNE (Fig. 2), researchers found that this cluster includes key skills such as mathematics, science, investment, analytics, management, architecture, logistics, and finance. The presence of companies like Oracle, Twitter, Panasonic, Tesla, and McDonald's suggests frequent job postings in this field.

Additionally, using the results from UMAP (Fig. 1), researchers found that this cluster is near **Cluster 7: Travel and Legalities**, suggesting the presence of jobs related to analyzing the legal aspects of finance and technology, such as financial advisors or RegTech specialists. It is also near **Cluster 8: Production** which suggests relevance of jobs such as Financial Data Engineer or Financial Supply chain analysts.

This cluster is also closely related to **Cluster 1: Health and Finance**, indicating jobs at the intersection of technology and healthcare, such as biomedical engineers or healthcare IT specialists.

Also, its proximity to **Cluster 6: Legal and relations** suggests the existence of tech-related law jobs, such as Data Privacy Attorney or Regulatory Compliance Officers. Lastly, **Cluster 0** is near **Cluster 5: Banking** which suggests existence ofo jobs such as FinTech or digital banking specialist.

Notably, technology is positioned near the center of the UMAP plot (Fig. 1) and is scattered throughout the graph, implying that a significant portion of LinkedIn job postings are related to technology in some capacity.

**Cluster 1: Health and Finance**

Based on the TSNE results (Fig. 2), the researchers found that this cluster emphasizes skills such as planning and compliance, highlighting their importance in these industries. Companies such as Pfizer, Sony, and Chevron frequently post job openings related to health and finance. Additionally, biotechnology and real estate appear prominently, linking technology and finance jobs to these sectors. The presence of *student loan assistance* suggests that jobs in health and finance often offer this benefit.

Using the UMAP results (Fig. 1), the researchers found that **Cluster 2** is closely related to **Cluster 7: Legal**, suggesting the existence of jobs at the intersection of health, finance, and law, such as healthcare insurance and risk management positions. Furthermore, **Cluster 1** is near **Cluster 4: Engineering and Design**, which suggests existence of role like Health care data analyst or health care financial analyst. Lastly, its proximity to **Cluster 3: Data and Security** suggests the relevance of roles that integrate medicine, finance, data, and security, such as healthcare data analysts or cybersecurity analysts in the healthcare sector.

Notably, **Cluster 1** is positioned near the center of the graph, indicating that many job postings on LinkedIn are closely related to medicine or finance.

**Cluster 2: Medicine, Environment, and Safety**

Using the findings from TSNE (Fig. 2), this cluster includes keywords such as *food* and *agriculture*, suggesting that many jobs in this field are related to these areas. Additionally, companies like KFC, Netflix, Domino's, and PepsiCo frequently post job openings within this industry. The positions being hired may involve *food quality assurance* or *food safety maintenance*, aligning with the cluster's focus on medicine, environment, and safety.

Based on the results from UMAP (Fig. 1), this cluster is closely related to **Cluster 5: Banking**, suggesting the presence of roles around health insurances. **Cluster 2** is also near **Cluster 8: Production**, indicating the existence of jobs related to agriculture such as food production or farming. Additionally, its proximity to **Cluster 6: Legal and Relations** suggests the relevance of Health care lawyers.

The distribution of **Cluster 2** is notably scattered near the edges of the graph, suggesting that professions related to environment and safety are highly specialized and inflexible.

**Cluster 3: Data and Security**

**Cluster 3** contains words such as entertainment, sales, consulting, regulatory, manufacturing, advertisement and supply chain which suggests that job postings in linked in revolves around data and security handling within these areas. Companies such as meta, adobe were also present which may be due to the artificial intelligence surge, requiring both companies to gather data related specialists to be ahead of the AI race. Cisco is also present which is expected for its security courses. An interesting result is the ups company which turns out to be a logistics company and logistics is nothing without data and security.

Additionally, **Cluster 4** is mapped near a small section of **Cluster 7: Travel and Legalities**, suggesting the existence of roles that intersect both fields, such as *data privacy lawyers* and *digital forensic analysts*. Its proximity to multiple clusters further suggests a high volume of job postings related to data security and cybersecurity.

**Cluster 4: Engineering and Design** The TSNE results (Fig. 2) for **Cluster 3** include keywords such as *finance*, *development*, and *safety*, suggesting that jobs related to Engineering and Design are primarily focused on finance and technological development. Companies such as Ferrari, Samsung, Microsoft, PayPal, and Google are present in this cluster, indicating that many of their job postings are related to Engineering an design which makes sense as these companies are known for technological innovations in engineering and design.

UMAP (Fig. 1) shows that **Cluster 4** is positioned near **Cluster 5: Banking** and is also related to a portion of **Cluster 9: Undefined**, suggesting the existence of various engineering-related jobs that the researchers were unable to clearly categorize.

**Cluster 5: Banking**

The results from TSNE (Fig. 2) indicate the presence of keywords such as *technology* and *branding*. This suggests that most engineering job postings on LinkedIn focus on technology development, while most design-related job listings pertain to brand creation. The company rbc was also inside which is a financial service company.

In addition to these connections, UMAP (Fig. 1) shows that **Cluster 5** is also related to portions of **Cluster 6: Legal and Relations** and **Cluster 7: Travel and Legalities**, both of which are expected associations. This suggests the presence of job roles such as *bank tellers* or *retail bankers*.

**Cluster 6: Legal and Relations**

Based on TSNE (Fig. 2), this cluster contains keywords such as *business, law, information, travel*, and *insurance*, suggesting that most job postings within this category pertain to business legalities and information management. Additionally, the presence of BPI, a banking institution, indicates that it is hiring for positions related to business and law.

Beside the relations already stated before, (fig 2) shows that it's related to **cluster 7: Travel and Legalities** which is expected **cluster 9: Production** which might suggest existence of jobs such as Corporate Lawyers or Contract Lawyers.

**Cluster 7: Travel and Legalities** This cluster contains mostly companies such as jpmorgan, boeing, facebook, airbus and unilever. Jp morgan is a financial institution, and both airbus an boeing are travel companies. This suggests that these companies has been hiring Travel or Legal professionals for the last years. The neighbors of this cluster has already been stated above.

**Cluster 8: Production** This cluster contains words such as motor, construction, food and farming which means that most job postings about production in linked in revolves around production of motor, construction of goods, production of food an farming. Additionally the company ford can be seen, suggesting that it has been Job hunting for applicants to help in its motor production.

All related clusters were already discussed previously

**Cluster 9: Undefined** The result of LDA for topic 9 was too over the place to produce insight upon.

## V. Comparison to other clustering algorithms

Again the researchers compared different clustering algorithms using word2vec an UMAP and gauged which one is the best using 3 different metrics. The results of the best clustering algorithm was then compared with LDA + TSNE and UMAP to gauge which produces better insights in terms of usefulness and realism.

| Metric | K means | DBSCAN | Agg | SC | MG |
|---|---|---|---|---|---|
| SC | 0.3802 | -0.3448 | 0.3080 | 0.3266 | 0.3686 |
| DBI | 0.7820 | 1.6276 | 0.8990 | 0.7268 | 0.7930 |
| CHI | 14954.97 | 228.8278 | 11995.687 | 9591.86 | 14269.33 |

TABLE I

Clustering Evaluation Metrics. SC = Silhoutte Score, DBI = Davies-Bouldin Index, CHI = Calinski-Harabasz Index, Agg = Agglomerative, MG = Mixture of Gaussians

K means was the best across all metrics, Mixure of Gaussians, followed by agglomerative, then Spectral Clustering and finally DBSCAN.

**Results and Insights of K Means**

Similar to LDA + UMAP and TSNE, Word2Vec + UMAP + K means similarly grouped relevant skills on specific industries. For example, K means similarly identified mathematics, science, production and manufacturing as relevant skills under technology, engineering, computer, architecture and education. It also similarly related legal with law, affairs and investment. Sales and strategy skills was also grouped under the business

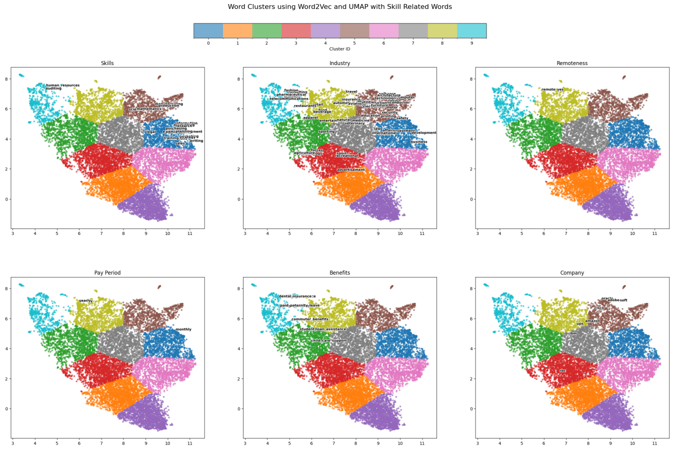

Fig. 3. Word2Vec + Umap + Kmeans

industry. Companies were also correctly mapped such as oracle, adobe and microsoft belonging to the brown cluster which is composed of tech-related industries. Unfortunately it was not able to map rbc correctly and placed it on the red cluster which delved around sports when it should have been in the blue cluster where the finance and business related skills were located.

Most importantly, Word2Vec + UMAP + K means was not able to properly identify the hierarchy between the words, hence mapping them in almost similar spots. This made the creation of insights between which industry are correlate to each other based on cluster distance almost impossible because some clusters contained no labels.

This result may be due to the nature of Word2Vec as it only excels in learning context up to a certain window allowing it to learn the grammatical relationship between words but misses the semantic meaning of a document most of the time.

## VI. Conclusion

The objective of this research is to streamline the process of navigating job postings by identifying key skills required for specific positions, determining the most relevant job opportunities, and understanding the relationships between different job sectors. Given the vast number of job listings available online, job seekers often face challenges in filtering through opportunities that align with their expertise and career aspirations. This study aims to provide a structured approach to categorizing job postings, highlighting essential skills, and uncovering connections between various industries. While the researchers have made significant progress in addressing these objectives, further refinement and analysis are required to enhance the accuracy and applicability of the findings.

The researchers found that Industries surrounding Business and Technology are closely related to Travel, Legalities, Production, Health and finance, hence suggesting that jobs like Financial Data Engineer, Financial Advisors, RegTech specialists, Supply Chain Analysts and Biomedical Engineers are relevant. Jobs on this field also required skills in areas such as mathematics, science, investment, analytics, management, architecture, logistics an finance. The researchers suggests pursuing technology as it provides a lot of flexibility in the job market. Companies under these includes Oracle, Twitter, Panasonic, Tesla an Mcdonald's

Next, the researchers found that Medicine, Environment and Safety relates with Banking, production and legal relations, suggesting the relevance of Jobs like Health care lawyers, Health insurance facilitators and food production. These are less relevant topics though and offerrs less flexibility in the job market. Companies along these are Pfizer, Sony and Chevron.

The researchers also found that data and security is also relevant among the fields of Travel an Legalities and travel. These opens the jobs for data privacy lawyers, digital forensic analysts. Companies among these are Meta and Cisco.

These insights will assist recruiters in identifying key areas for hiring priorities. Additionally, upcoming graduates can leverage these findings to make informed career decisions. More importantly, job seekers can utilize this data to guide their job search strategy before navigating the complex job market on LinkedIn.

Due to time constraints and limited manpower, the researchers acknowledge certain limitations in this study. While insightful, more advanced topic modeling techniques such as **BERTopic, Top2Vec, and the Embedded Topic Model (ETM)** could have provided enhanced results. However, these models were not implemented due to time constraints.

Furthermore, only two visualization techniques were employed in this study. Future researchers could refine the analysis by incorporating alternative **dimensionality reduction techniques**, such as **Autoencoders**, which may reveal deeper patterns within the data.

In comparing clustering algorithms, a more effective **word embedding technique** would have been **BERT**, as it captures semantic meaning more accurately than **Word2Vec**. However, similar to LDA, implementing **BERT** requires substantial computational resources and time, which were beyond the scope of this study. Additionally, the researchers did not apply rigorous metrics to validate the insights obtained. Future studies could improve upon this by introducing models that assess the reliability of each claim.

This research did not address crucial questions such as identifying industry-specific tools, especially in the tech sector, distinguishing between emerging and traditional job roles, determining the true clustering of job postings, or analyzing salary distributions within each cluster.

Overall, the researchers consider the findings valuable, particularly for individuals seeking to understand the relationships between different job industries.

## References

[1] BeBusinessEd, "History of the Online Job Search," July 2023.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2013.

[3] J. Shen et al., "Learning to Retrieve for Job Matching," unpublished, Feb. 2024. [Online]

[4] Indeed Engineering Blog, ”How Indeed Replaced its CI Platform with GitLab CI,” 2024. [Online]

[5] B. Shi, J. Yang, F. Guo, and Q. He, “Salience and Market-Aware Skill Extraction for Job Targeting,” in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, Aug. 2020, pp. 3026–3034. doi: 10.1145/3394486.3403251.

[6] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, ”A Neural Probabilistic Language Model,” *J. Mach. Learn. Res.*, vol. 3, pp. 1137-1155, 2003.

[7] A. Konstantinov, V. Moshkin, and N. Yarushkina, “Approach to the Use of Language Models BERT and Word2Vec in Sentiment Analysis of Social Network Texts,” in *Recent Research in Control Engineering and Decision Making*, vol. 186, Cham, Switzerland: Springer, Jan. 2021, pp. 462–473. doi: 10.1007/978-3-030-65283-8_38.

[8] B. Wang, A. Wang, F. Chen, Y. Wang, and C. J. Kuo, ”Evaluating Word Embedding Models: Methods and Experimental Results,” *APSIPA Trans. Signal Inf. Process.*, vol. 9, pp. 1-13, 2020.

[9] GeeksforGeeks, ”Spectral Clustering in Machine Learning,”

[10] Ann. Data. Sci. (2015) 2(2):165–193, DOI 10.1007/s40745-015-0040-1, A Comprehensive Survey of Clustering Algorithms, Dongkuan Xu1,2 · Yingjie Tian2,3.

[11] N. N. Narisetty, ”Bayesian model selection for high-dimensional data,” in *Handbook of Statistics*, vol. 43, Elsevier, 2020, ch. 4, pp. 207-248.

[12] L. van der Maaten and G. Hinton, ”Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[13] H. Li, J. Cui, X. Zhang, Y. Han, and L. Cao, ”Dimensionality reduction and classification of hyperspectral remote sensing image feature extraction,” *Remote Sensing*, vol. 14, no. 18, p. 4579, Sep. 2022.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, ”Latent Dirichlet Allocation,” *J. Machine Learn. Res.*, vol. 3, pp. 993–1022, 2003.