# Fake Job Postings Detection

Ursu Andrei
andrei.ursu@s.unibuc.ro

January 27, 2025

**Abstract**

This project proposes several unsupervised machine learning models with the aim of determining whether a job posting is fake or real. The dataset contains a column for text and a column for the label, 0 for real job postings and 1 for fake job postings. The best models for this task

With the rise of digital recruitment platforms, job seekers increasingly rely on online job postings to find employment opportunities. However, this growth has also attracted malicious actors who post fraudulent job advertisements to exploit unsuspecting individuals. These fake job postings can lead to identity theft, financial fraud, and other harmful consequences. Detecting such fraudulent postings is a critical task to ensure a safe online job market.

## 1 Methodology

In this section I will present the preprocessing method that obtained the best results as well as a short description for each model that I used. For preprocessing, I chose to keep the text exactly as it is and only vectorize it, because from my experiments, I observed that removing stop words and punctuation negatively affects the results.

### 1.1 Dataset

The dataset contains 7590 job postings, of which 3795 are real and 3795 are fake. The original dataset includes several columns, such as job title, job requirements, industry, benefits and salary range; however, we will only use the job description column, which contains a brief description of the job, and the fraudulent column, which indicates whether the posting is real or fake. The following graphs illustrate the length measured in both words and letters for the real inputs and the false ones.

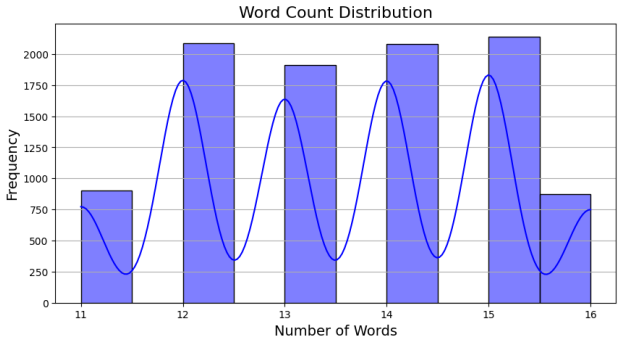


Figure 1: Fake Word Count

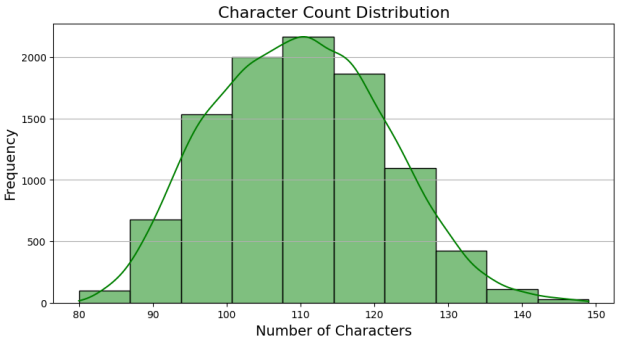


Figure 2: Fake Chars Count

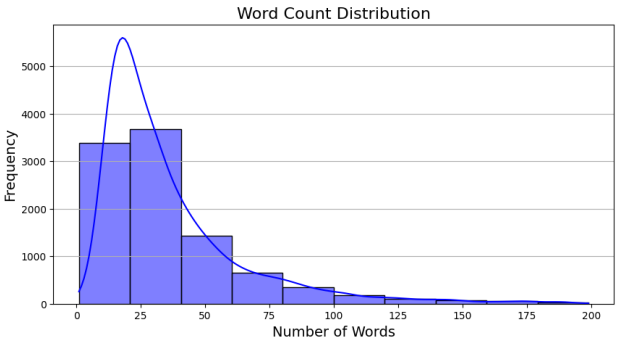


Figure 3: Real Word Count

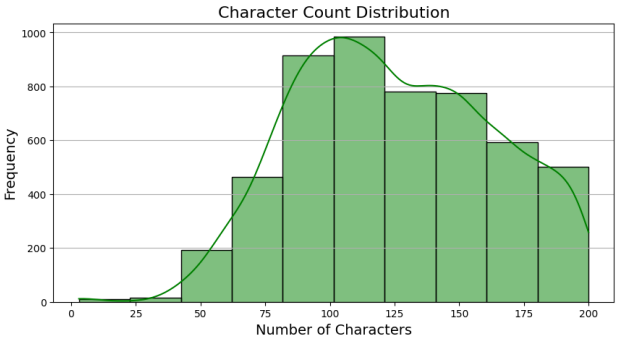


Figure 4: Real Chars Count

## 1.2 TF-IDF (Term Frequency-Inverse Document Frequency) Representation

TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (the corpus). It is a popular method for transforming text data into numerical features.

**How it works:**

- **Term Frequency (TF):** Measures how often a term appears in a document.
- **Inverse Document Frequency (IDF):** Reduces the weight of commonly occurring words (like "the" or "is") that appear in many documents, giving more importance to rare words.

The formula for a word $t$ in a document $d$ is:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \log\left(\frac{N}{1 + \text{DF}(t)}\right)$$

where $N$ is the total number of documents and $\text{DF}(t)$ is the number of documents containing the term $t$.

**Why use it?**

- It creates sparse, high-dimensional vectors for text, highlighting the most important words while suppressing irrelevant ones.
- It works well with bag-of-words approaches and is suitable for algorithms like SVM, logistic regression, and clustering methods like Affinity Propagation.

**Limitations:**

- It does not capture word semantics or context (e.g., synonyms or polysemy).
- High-dimensional output can increase computational complexity.

### 1.3 Word Embeddings with GloVe

GloVe (Global Vectors for Word Representation) is a pre-trained word embedding model that encodes semantic relationships between words into dense, low-dimensional vectors.

**How it works:**

- GloVe learns word vectors based on the **co-occurrence matrix** of words in a large corpus.
- It optimizes a cost function that ensures words with similar contexts (i.e., co-occurring with similar words) are mapped to nearby vectors in the embedding space.
- Each word is represented by a dense vector of a fixed size (e.g., 50, 100, 200 dimensions), where dimensions capture latent linguistic features like syntax and semantics.

**Why use it?**

- Captures both semantic similarity (e.g., "king" is close to "queen") and analogies (e.g., "king - man + woman = queen").
- Pre-trained on large corpora, saving computation time and improving performance on small datasets.

**Limitations:**

- Cannot handle out-of-vocabulary words (words not in the pre-trained corpus).
- Sentence-level embeddings (as in your code, averaging word embeddings) might oversimplify by ignoring word order and syntactic structure.

### 1.4 Methodological Approaches

Our analysis employs various machine learning techniques, beginning with classical approaches to establish baseline performance.

### One-Class SVM

The One-Class SVM (Support Vector Machine) is an unsupervised learning algorithm primarily used for anomaly detection. It works by identifying a decision boundary that encapsulates the majority of the data points in the feature space, treating them as "normal." Points falling outside this boundary are classified as anomalies or outliers. The algorithm maps data into a higher-dimensional space using a kernel function (e.g., linear, RBF), allowing it to learn complex boundaries. It is particularly useful when only "normal" data is available during training.

### Affinity Propagation

Affinity Propagation is a clustering algorithm that identifies clusters by exchanging messages between data points. Instead of pre-specifying the number of clusters, it determines them based on data similarities. Each point can serve as an "exemplar" (a cluster center), and the algorithm iteratively refines clusters by optimizing two types of messages: responsibility (how suitable a point is to be an exemplar) and availability (how appropriate it is for a point to belong to an exemplar). This process converges to a set of exemplars and their corresponding clusters, making it effective for discovering patterns in data with varying cluster shapes and sizes.

### Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and aggregates their predictions to improve accuracy and reduce overfitting. By randomly selecting subsets of data and features for each tree, it ensures diversity among the trees, which leads to a more robust model. For classification, it uses majority voting to make predictions, and for regression, it averages the results of the trees. The algorithm is effective for handling large datasets with both numerical and categorical features, but it can be computationally intensive and harder to interpret compared to a single decision tree.

## Experimental Results

This section presents an analysis of our experimental outcomes, highlighting both successful and unsuccessful approaches. For Linear Regression, Random Forest Regressor, and Decision Tree Regressor models, we use **Spearman's Rank Correlation** as our primary **score** metric. The Gradient Boosting Regressor, XGBoost Regressor, and K-NN models were evaluated using the **6000 features** preprocessing pipeline.

### Evaluation Metrics

For One-Class SVM, the metric used was accuracy, while for Affinity Propagation, the metric used was the Adjusted Rand Index. The Adjusted Rand Index (ARI) is a measure of the similarity between two clusterings, accounting for chance. It is defined as:

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}},$$

where:

- $n$ is the total number of samples,

- $n_{ij}$ is the number of samples in both cluster $i$ of the ground truth and cluster $j$ of the predicted clustering,
- $a_i = \sum_j n_{ij}$ is the number of samples in cluster $i$ of the ground truth,
- $b_j = \sum_i n_{ij}$ is the number of samples in cluster $j$ of the predicted clustering.

The ARI ranges from $-1$ to 1, where:

- 1 indicates perfect agreement between the clusterings,
- 0 indicates a clustering assignment no better than random chance,
- Negative values indicate disagreement between the clusterings.

### 1.5 One-Class SVM

For One-Class SVM, the best results were obtained using the "rbf" kernel and "nu=0.01", where the accuracy was 0.5019. Below, we will present some of the best results using the "rbf" kernel and a "nu" value varying between 0.01 and 0.03. Also, the first figure shows the real clustering for comparison.

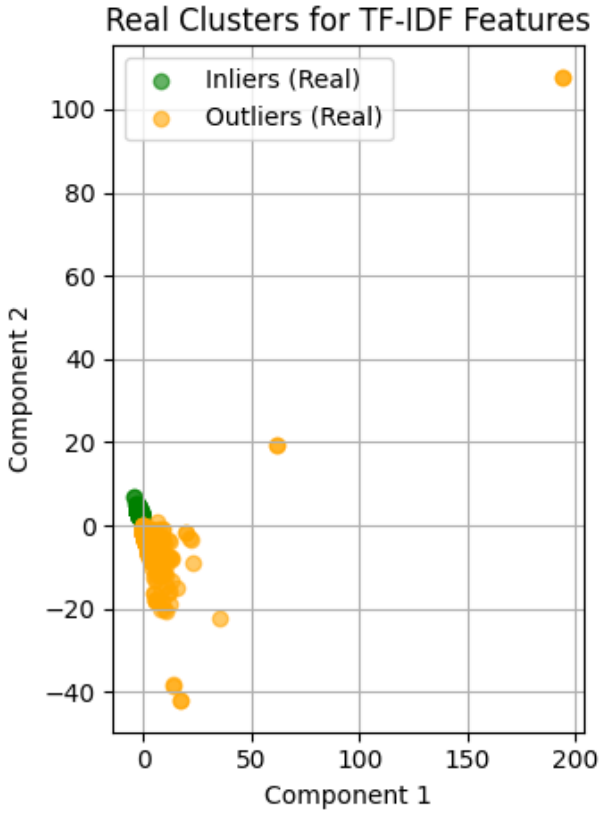


Figure 5: tfidf, real distribution

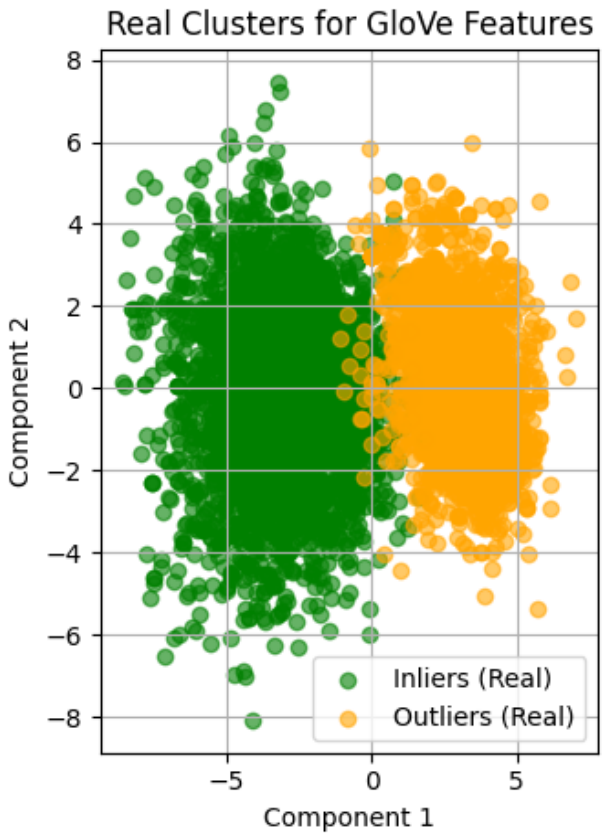


Figure 6: glove, real distribution

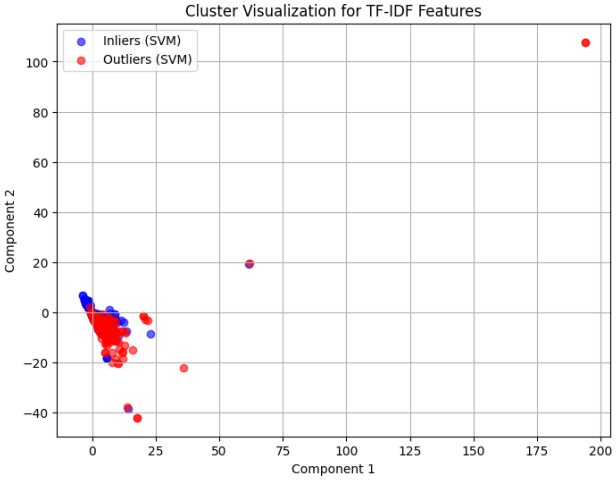


Figure 7: SVM, tfidf, nu=0.01

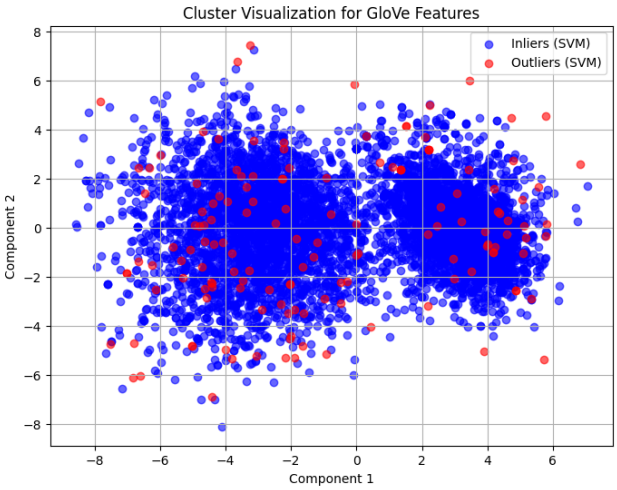


Figure 8: SVM, glove, nu=0.01

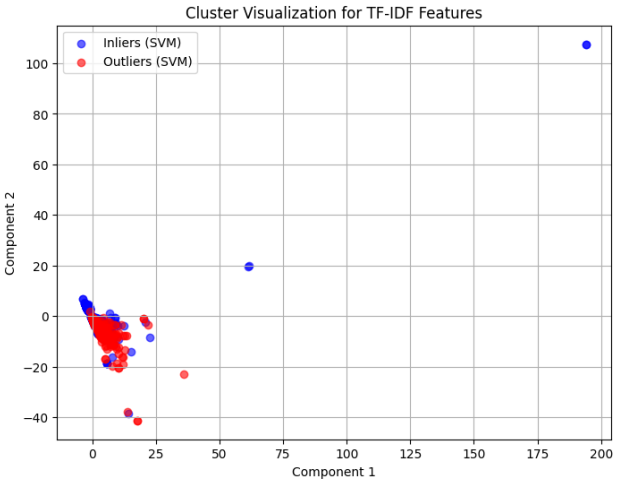


Figure 9: SVM, tfidf, nu=0.02

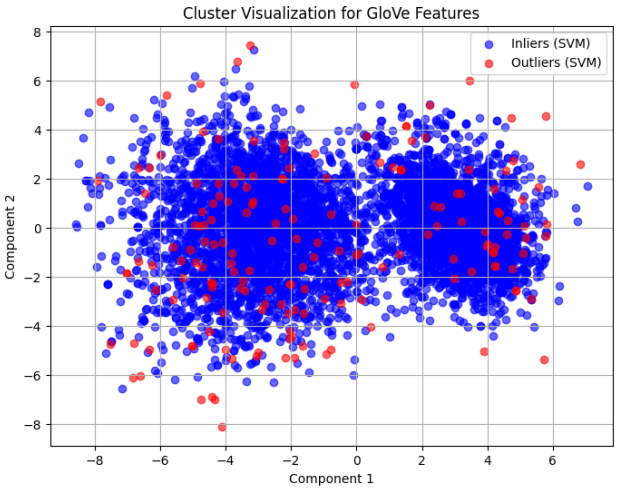


Figure 10: SVM, glove, nu=0.02

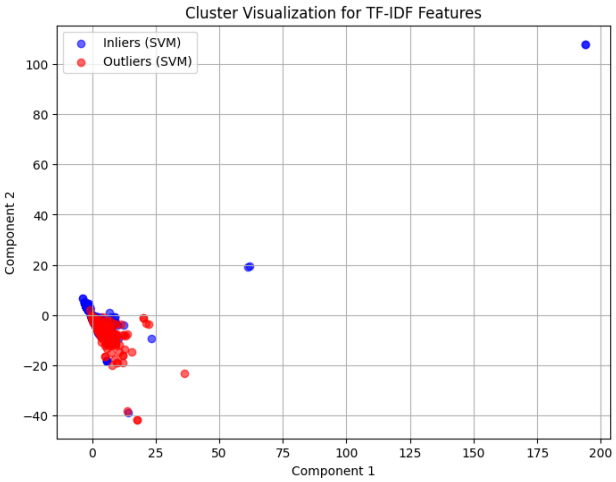


Figure 11: SVM, tfidf, nu=0.03

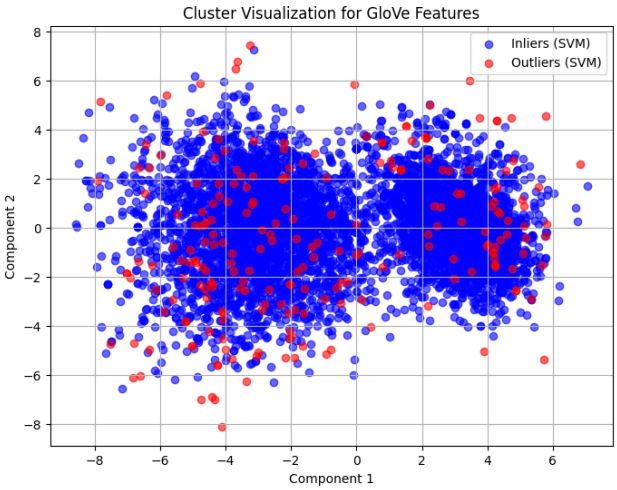


Figure 12: SVM, glove, nu=0.03

## 1.6 Affinity Propagation

For Affinity Propagation, the results were not very good. The best model used "damping = 0.5" and "preference = -20," achieving an ARI score of only 0.4. Below, we present the results using "preference = -20" and a damping value varying between 0.5 and 0.7. Additionally, the first figure shows the ground truth clustering.

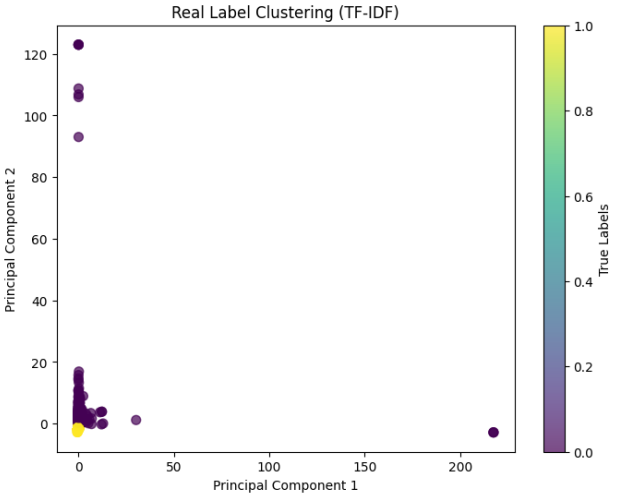


Figure 13: Affinity Propagation, real clustering

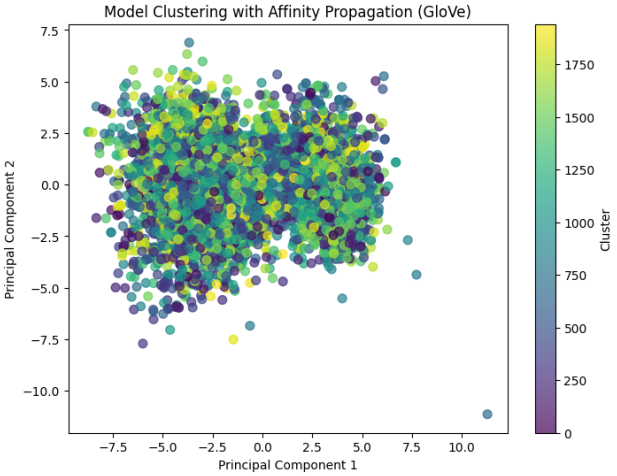


Figure 15: Affinity Propagation, glove, damping=0.5

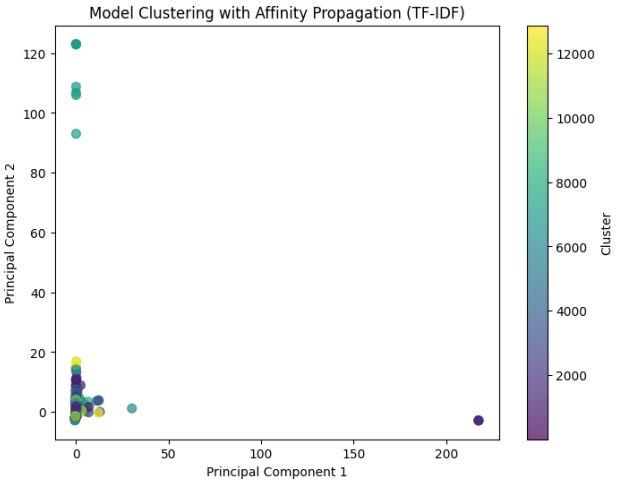


Figure 14: Affinity Propagation, tfidf, damping=0.5

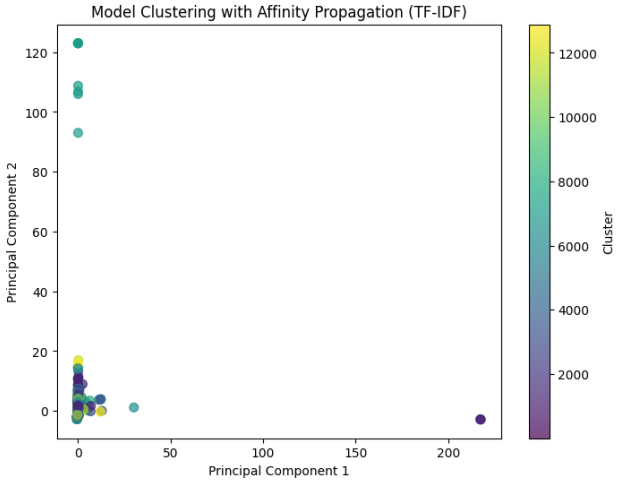


Figure 16: Affinity Propagation, tfidf, damping=0.6

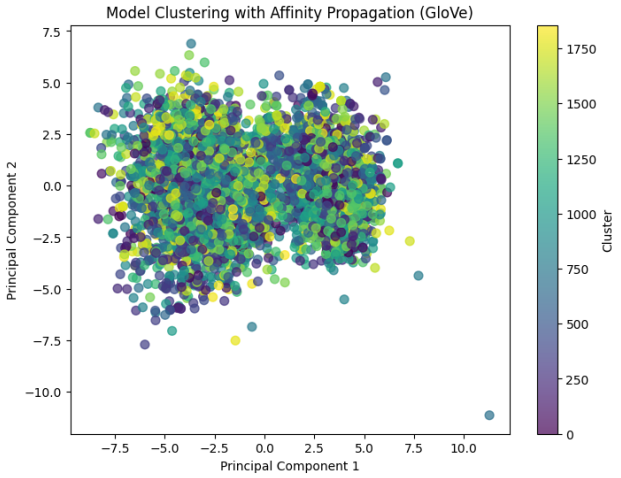


Figure 17: Affinity Propagation, glove, damping=0.6

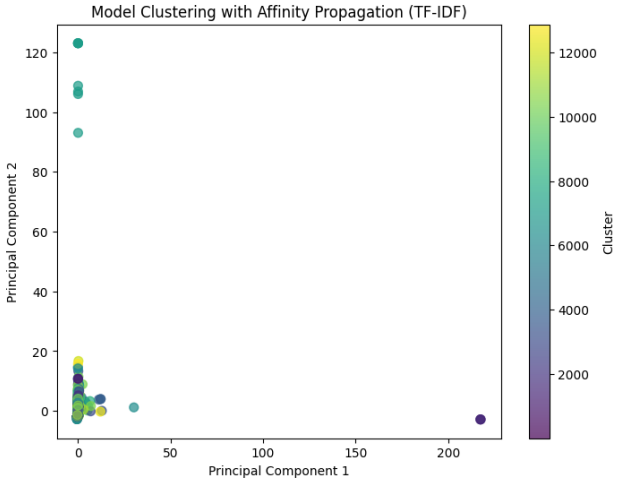


Figure 18: Affinity Propagation, tfidf, damping=0.7

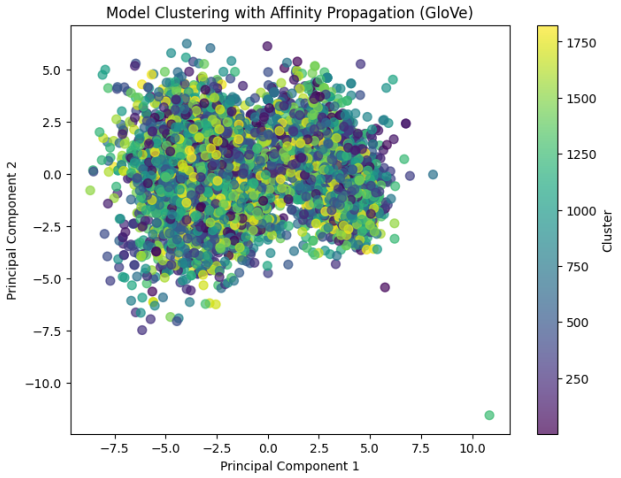


Figure 19: Affinity Propagation, glove, damping=0.7

## 1.7 Random Forest

As expected, the results obtained with Random Forest were the best. For both types of data preprocessing, we achieved an accuracy of approximately 99%. Below, we present a comparison of the supervised algorithm with One-Class SVM. For reference, we also included random chance on the right, which was set at 0.5.

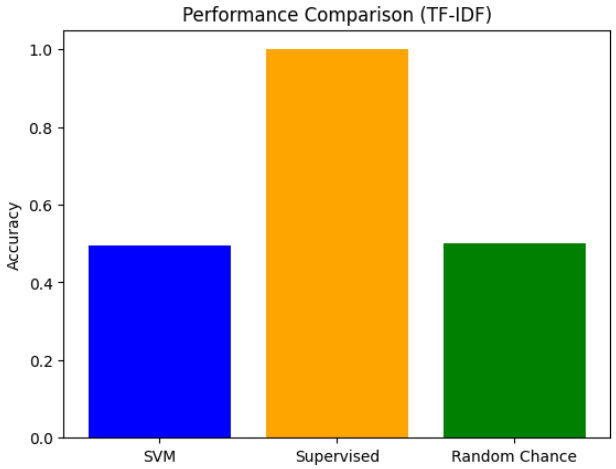


Figure 20: SVM-tfidf-nu=0.01 vs RF vs Random Chance

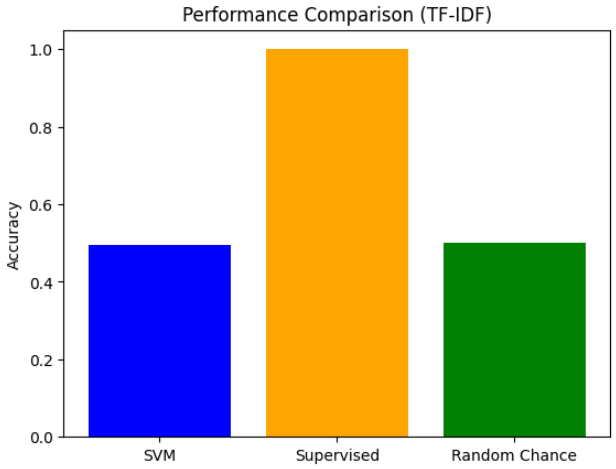


Figure 21: SVM-tfidf-nu=0.02 vs RF vs Random Chance

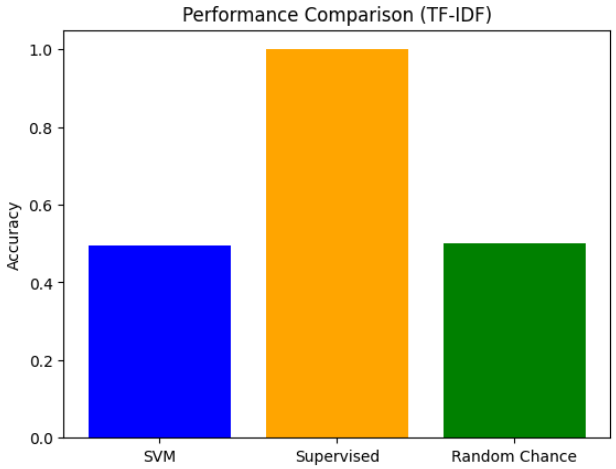


Figure 22: SVM-tfidf-nu=0.03 vs RF vs Random Chance

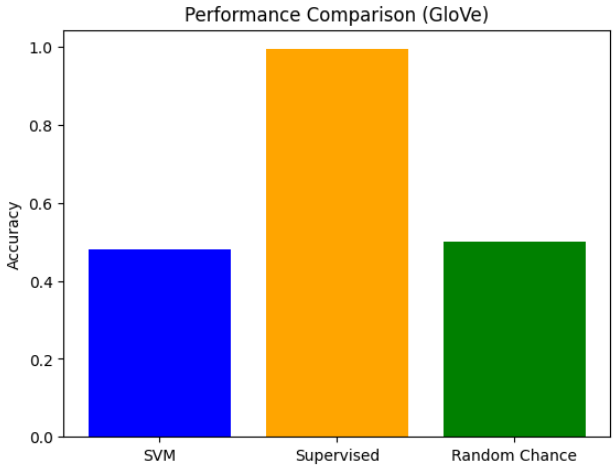


Figure 23: SVM-glove-nu=0.01 vs RF vs Random Chance

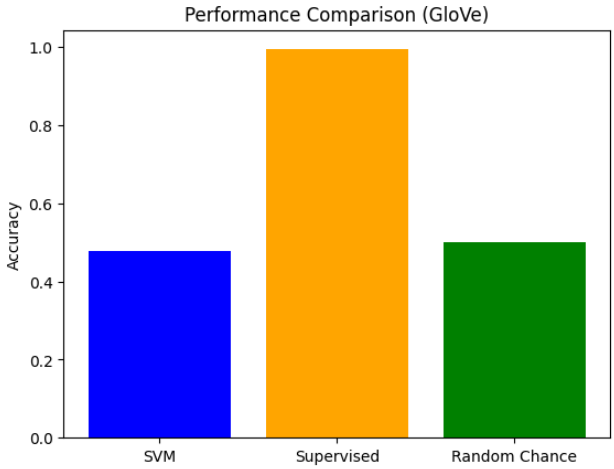


Figure 24: SVM-glove-nu=0.02 vs RF vs Random Chance

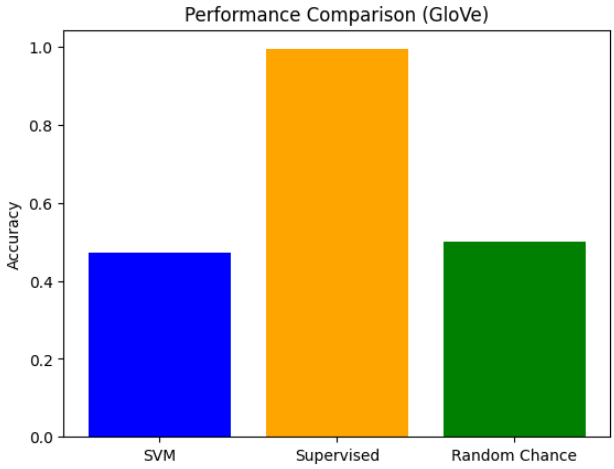


Figure 25: SVM-glove-nu=0.03 vs RF vs Random Chance

## 1.8 K-means

Finally, to compare all the results obtained with a more versatile clustering model, we used K-Means, where ARI was also used as the evaluation metric. The results obtained with both preprocessing methods are significantly superior compared to the other two unsupervised models. For TF-IDF, we achieved an ARI of 0.9736, while for GloVe, we obtained an ARI of 0.9108. Below, the figures illustrate the clusters obtained.

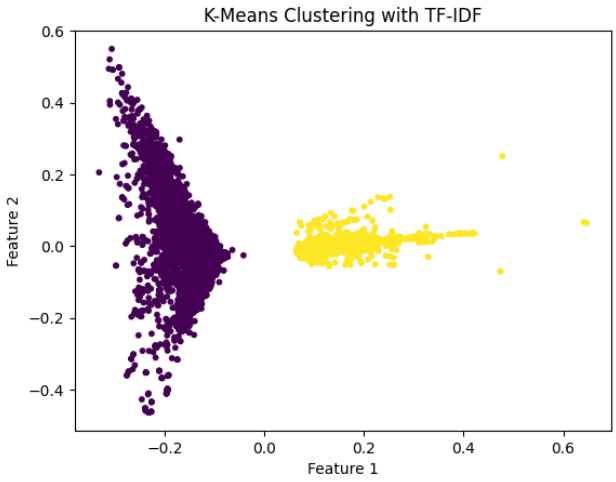


Figure 26: K-means tf-idf

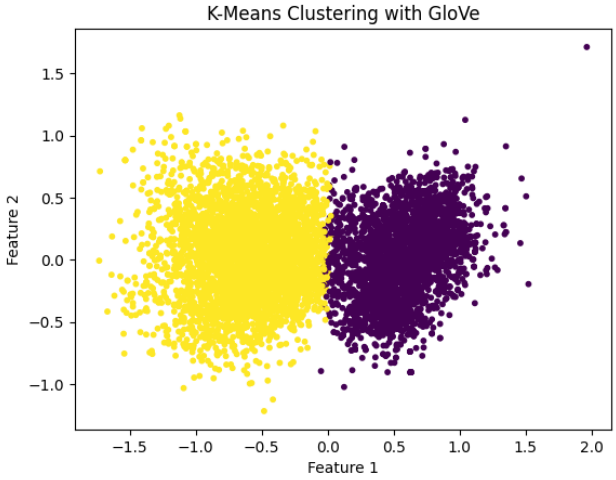


Figure 27: K-means GloVe

## 1.9 Error Analysis

For One-Class SVM, some of the most misclassified for fake entries for all configurations were:

- Saleswoman will sell menswear and accessories.
- The United Nations Development Programme in Armenia announces opening for the position of Country Economist.
- Basic knowledge in front, no degree required. Flexible hours.
- AMERIA Closed Joint Stock Company is seeking a Receptionist to provide secretarial and administrative support to the office.
- The company WEB PROJECT is currently seeking qualified candidates for positions of associates in work-online business.

- ACH's Armenia Mission is seeking to employ a Administrative Assistant/ Secretary for the Sisian Base office.
- - 6 days working week - Company provides new cars (Zhiguly 06), with an opportunity to acquire it in future.
- Excellent knowledge of Accounting/Tax filing both Central Bank and Tax Dept., budget formation, presentation and control.
- The UNDP and the Ministry of Health seek professionals for the project "HIV/AIDS and Uniformed Services"
- Yerevan Brandy Company is seeking qualified candidates to fill the position of Technical Project Manager.

For Affinity Propagation, some of the most misclassified for fake entries for all configurations were:

- Valensia Hotel & Resort is looking for a Marketing Specialist.
- Antares Media Holding is looking for an experienced Artist - Designer.
- Saleswoman will sell menswear and accessories.
- The appointment is in replacement capacity for one year GL-5
- Yerevan Brandy Company is seeking qualified candidates to fill the position of Technical Project Manager.
- The individual will take part in design, implementation, and execution of software tools.
- A qualified programmer is needed in order to participate in projects.
- Excellent knowledge of Accounting/Tax filing both Central Bank and Tax Dept., budget formation, presentation and control.
- The company WEB PROJECT is currently seeking qualified candidates for positions of associates in work-online business.
- Distribution of cosmetic and laundry products to retail points in Yerevan.

## 2 Conclusion

In conclusion, this project has demonstrated that for this binary classification task, the best method turned out to be the most commonly used one. Although Affinity Propagation achieves impressive results in various scenarios, it did not prove to be as effective for this particular problem.