2022 30th International Conference on Electrical Engineering

Title: Job Title Prediction from Tweets Using Word Embedding and Deep Neural Networks

Presented by: Shayan Vassef

Authors with Affiliation:

Shayan Vassef,Ramin Toosi, Mohammad Ali Akhaee

*School of Electrical and Computer Engineering University of Tehran Tehran, Iran

†Adak Vira Iranian Rahjo company Tehran, Iran

Topics

Job Title Prediction

1. Motivation
2. Related Work
3. Data Gathering
4. Proposed Method
5. Evaluation

Motivation

- Social media had become quite popular.
- People’s job affect their activities on social media.
- Proposing a novel method to predict the **job title** on **Twitter**.
- ***Job*** is a highly semantic target.
- Create a dataset from scratch.

Related Work

- One of the most important research topics in social media analysis is **event detection** approaches.
- Removing irrelevant tweets on Twitter by Harris et al.
- Implementing a text classification algorithm on landslide detection by Musaev et al.
- Detecting health organizations 'tweets during the Covid-19 pandemic by Petersen et al.
- Appling a probabilistic model (HDP) to cluster tweets with similar trending topics by Madani et al.
- Integrating both textual and imagery content to improve the precision of event detection by Hou et al.

Data Gathering

To create our dataset, we need to pass through two essential steps:

- **Discovering celebrities:** Designing an algorithm to form a dataset of Celebrities with their Twitter accounts.
- **Job search:** Designing an algorithm to extract jobs by crawling the Wikipedia webpages.

Data Gathering

We designed two approaches for Data Gathering:

- **Search by emoji:** creating a list of selected emojis utilized by the users on Twitter.
- **Search by hashtags:** Same procedure, replacing hashtags with emojis this time.

Our raw dataset consisted of the users' names and usernames by running the two above algorithms.

What is the benefit of having users' names?

Data Gathering

Next, we extracted the corresponding bio and tweets for each username :

Three limits were set for the tweets:

- All tweets must be written in **English**.
- Each user should have a minimum number of **10** and a maximum number of **40** tweets.
- Set a boundary of **50K** on the minimum number of likes for the tweets.

Data Gathering

why we focused on the celebrities community ?

Example: ”Katheryn Elizabeth Hudson, known professionally as Katy Perry, is an American singer, songwriter, and television judge. After singing in church during her childhood, she pursued a career in gospel music as a teenager.”

The user’s jobs are <u>American singer</u>, <u>songwriter</u>, and <u>television judge</u>.

How to extract the information ?

The user’s job titles are explained in the first two sentences of the **infobox**.

*(Searching for a user is done by the user’s name.)

Each phrase related to job titles begins with the verb ”is” and ends with a dot(·).

Exceptions : When the user passes away or when searching for a group like music bands in which our phrase starts with the verb ”are.”

Data Gathering

- Rule-Based method
  - Cleaning
    - **Removing nationalities:** phrases like American-born or American-Italian-born .
    - **Removing useless words:** Examples of these words are ”professional” or ”retired.”
  - Reducing
    - **Essential lists:** ”known for” expression is an **Essential lists** in the ”Lila Diane Sawyer is an American television broadcast journalist known for anchoring major programs” sentence.
    - Our desire phrase is between the specific verb discussed earlier and the essential lists.
    - The reduced phrase will be American television broadcast journalist.

Proposed Method

Employ K-means clustering.

- Choosing **Glove pre-trained model** as an embedding model.
- Using glove.6B.100d, where a vector with 100 elements shows each word.

Preprocessing user information (tweets, bios, Etc.).

Training a Classification model.

Proposed Method

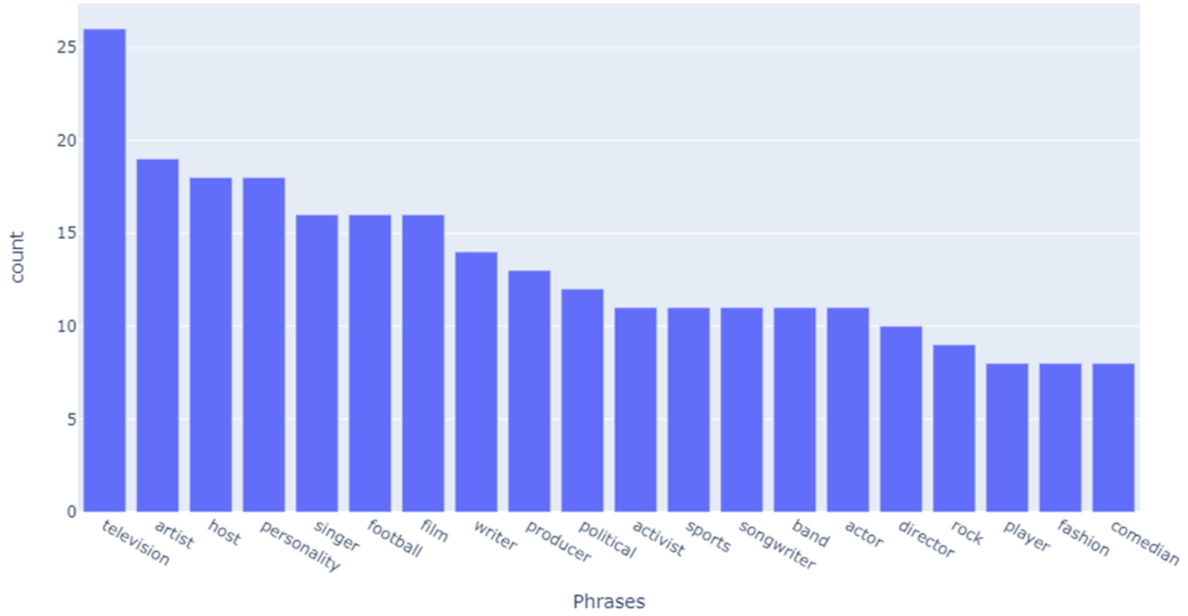


Fig. 1. the 20 most frequent unigrams appearing inside the job titles

Two significant challenges before the clustering process:

- Each individual has several occupations.
  - selecting the first two jobs of the users.
- Each job may have several words.
  - How can we convert that job to the simplified version without losing much information?

Proposed Method

Find the closest meanings:

- Ranking the first n words by closeness to a word.
- Using the combination of math operations compatible with vectors

Example:

- User's first two jobs are ”video game programmer” and ”designer.”
  - ”video game programmer” : ”video+ game+ programmer” in the new vector representation.
  - The word ”game” will be the closest equivalent word.
- “Television host,” “Television producer,” and “broadcast journalist” will convert to “Television”.

Proposed Method

Preparing for clustering:

The user’s jobs with two words should have a constant length vector representation similar to those with one word.	Apply Sum Word Vectors method to Equalize the length vector for each user’s jobs.	Considering the same weight for the first and second words: $(\textbf{\textit{The assigned vector to each user}} = 1 * vector(fisrt\ word) + 1 * vector(second\ word))$

Results:

A matrix with the dimension of 1314*100 is the input to the clustering process.	The optimum number of groups is **nine**, meaning there are nine job titles.

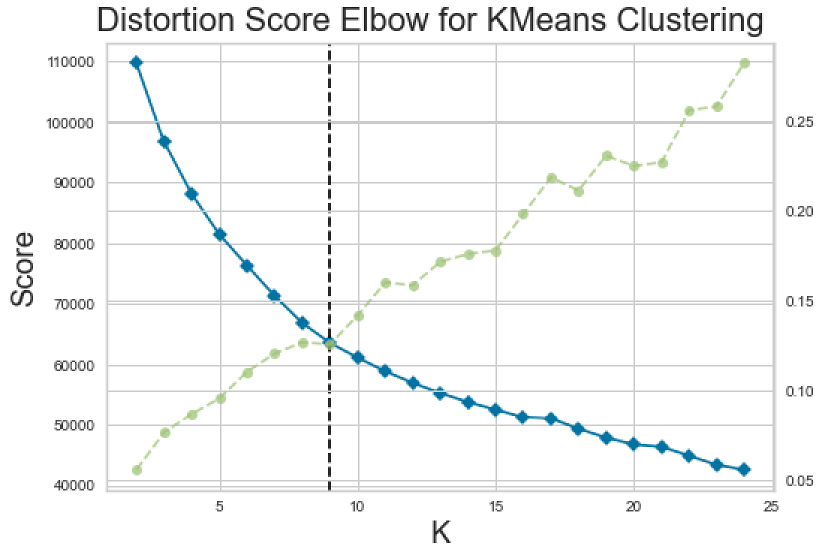


Fig. 2. The optimum number of clusters using the elbow method happens at k = 9.

Proposed Method

Finding patterns in each cluster:

The most repetitive pattern inside each cluster is demonstrated in **Table I**.

”mixed” means that various job titles exist with low relation in that group.

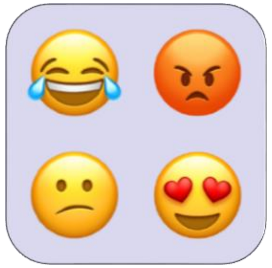
We will investigate how this condition affects the performance of the proposed method in labels 2 and 8 later.

TABLE I
PATTERNS FOR EACH LABEL

Label	***Pattern***
0	singe&songwriter
1	politician
2	singer&**mixed**
3	footballer
4	actor-actress&singer
5	basketball
6	actor-actress&comedian
7	rapper&singer
8	television&**mixed**

Proposed Method

There are two items inside a text that may enhance our performance:



Emojis:

- Emojis are often used to show emotions.
- Some emojis can be related to jobs.
  - There exist eight types of emojis: ”Smileys People,” Animals-Nature,” Food-Drink,” Activity,” Travel-Places,” Objects,” Symbols,” and ”Flags.”
  - The most related category to jobs is the ”Activity” type.

Hashtags:

- Three approaches in facing hashtags:
  - **Method 1(RHW):** Remove both the hashtag sign(#) and the following word.
  - **Method 2(RH):** Remove just the hashtag sign(#).
  - **Method 3(RHRW):** Remove the hashtag sign(#) and replace the following word with its most relevant string.

Proposed Method

- Investigate embedding techniques and classifiers:
  - Four popular embedding techniques, Bag of words, TF-IDF, Word2vec, and Glove
  - Three classifiers: logistic regression, deep neural network (DNN), and convolutional neural network (CNN).
    - Logistic regression:
      - implemented by multinomial class .
      - solved by the ’Newton-CG’ algorithm.
    - DNN:
      - We employed the TF-IDF method as the embedding model.
      - 512 nodes in each hidden layer.
      - Considering dropout for generalization purposes.
    - CNN:
      - We employed the Glove pre-trained model for the embedding process.
      - Three convolutional layers and a global max pooling at the end.
      - Passing through a fully connected part with 128 nodes.

Evaluation

The summary of models 'performance is depicted in Table II.

The model performance for each method would not differ when using simple models like logistic regression and DNN.

The result would worsen by making the model more complex (DNN to CNN).

”RH” outperformed in comparison with ”RHW” and ”RHRW” when we used CNN as a classifier.

On Average,” RH” and ”RHRW” worked similarly in each case.

- The bios and hashtags do not explicitly enhance the accuracy of the proposed model.

TABLE II
FINALL ACCURACIES FOR THE THREE METHODS

	Accuracy		
Model	***RHW***	***RH***	***RHRW***
Logistic	53.6	54	52.1
DNN	54	53	52
CNN	38	40	38

Evaluation

Per-class evaluation:

- In Table III, we have a classification report of the DNN model with 53% accuracy on the ”RH” method .
- The high performance of the third label with a precision of 76% was observed.
  - The third label shows the footballers group.
  - Labels 2 and 8 were below the average of the model performance, and the rest of the labels were on average.
  - Both were mixed groups where it was hard to assign a shared label or job title to them.

TABLE III
DNN REPORT FOR METHOD2(RH)

DNN classification report			
Label	***precision***	***recall***	***f1-score***
0	59	45	51
1	60	43	50
2	30	41	34
3	76	76	76
4	61	61	61
5	61	61	61
6	55	59	57
7	56	58	57
8	42	28	33
M-avg	56	52	53
W-avg	54	53	53

Thanks for your attention !