

# An Experience with Text Classification in Datadays 2019

Majid Hajiheidari   Amirmohammad Asadi

April, 2019

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

# Divar Posts Dataset

- ▶ Released for DataDays 2019
- ▶ One million posts

بارگشت همه آگهی ها / املاک / اجاره مسکونی (آپارتمان، خانه، زمین) / آپارتمان



**190 متر/4 خواب/فول محدوده کاج**  
دقایقی پیش

دریافت اطلاعات تماس

شروع چت

نشان کردن

دسته‌بندی	آپارتمان
محل	تهران سعادت‌آباد
نوع آگهی	ارائه
آگهی‌دهنده	شخصی
تعداد اتاق	چهار
متراژ	۱۹۰

بارگشت همه آگهی ها / سرگرمی و فراغت / دوچرخه/اسکیت/اسکوتر

بارگشت همه آگهی ها / سرگرمی و فراغت / دوچرخه/اسکیت/اسکوتر



**دوچرخه مریدا 7-300 BIG سال ۲۰۱۷**  
۲ ساعت پیش

دریافت اطلاعات تماس

شروع چت

نشان کردن

دسته‌بندی	دوچرخه/اسکیت/اسکوتر
محل	تهران میدان آزادی
نوع آگهی	فروشی
قیمت	۵,۸۰۰,۰۰۰ تومان

با سلام یک دستگاه دوچرخه مریدا 7-300 BIG سال ۲۰۱۷ در حد آک آک سایز 27/5 تنه 18/5 با کمک پاد ست لوزیم دنده=طبیق و

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

# Columns

- ▶ id
- ▶ archive\_by\_user
- ▶ published\_at
- ▶ **cat1**
- ▶ **cat2**
- ▶ **cat3**
- ▶ city
- ▶ **title**
- ▶ **desc**
- ▶ price
- ▶ image\_count
- ▶ platform
- ▶ mileage
- ▶ brand
- ▶ year
- ▶ type

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

# The Problem: Categorization

- ▶ We need to categorize posts based on other posts features;
- ▶ We only use text features(title & description)!

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

#### Features

No. of Classes

First approach:  
Naive Bayes

Count Vectorizer

Bayes Classifier

Hyperparameters

Second approach:  
CNN

Sequence of vectors

How We Embed

Patterns with CNN

Dense Classifier

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features

No. of Classes

First approach:  
Naive Bayes

Count Vectorizer

Bayes Classifier

Hyperparameters

Second approach:  
CNN

Sequence of vectors

How We Embed

Patterns with CNN

Dense Classifier

The End

# First Approach: Naive Bayes Classifier

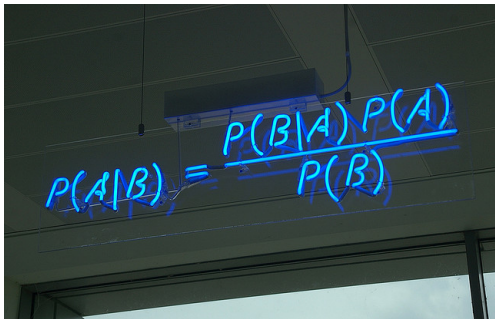

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Photo by Matt Buck



An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

# Vectorizing the Text: Count Vectorizer

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

An example: We want to vectorize these 4 sentences<sup>1</sup>:

1. Hello, how are you!
2. Win money, win from home.
3. Call me now
4. Hello, Call you tomorrow?

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

---

<sup>1</sup>Example from Rahul Vasaikar



# Vectorizing the Text: Count Vectorizer

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

1. We first build a vocabulary:

*vocabulary =*

*{are, call, from, hello, home, how, me, money, now, tomorrow, win, you}*

2. Then, we vectorize each sentence based on the occurrence of each word:

	are	call	from	hello	home	how	me	money	now	tom...	win	you
1	1	0	0	1	0	1	0	0	0	0	0	1
2	0	0	1	0	1	0	0	1	0	0	2	0
3	0	1	0	0	0	0	1	0	1	0	0	0
4	0	1	0	1	0	0	0	0	0	1	0	1

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

# Vectorizing the Text: Count Vectorizer

N pair of samples

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

**Count Vectorizer**  
Bayes Classifier  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

# Bayes Classifier: Naive One!

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

It is possible to show that accuracy is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values. In other words, we should simply assign a test observation with predictor vector  $\mathbf{x}_0$  to the class  $j$  for which

$$P(Y = j \mid \mathbf{X} = \mathbf{x})$$

is largest.

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
**Bayes Classifier**  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

# Bayes Classifier: Naive One!

We make two assumptions:

1.  $X_1, X_2, \dots$ , and  $X_m$  are independent from each other;
2.  $X_1, X_2, \dots, X_m \mid Y \sim MN(\cdot, p_1, p_2, \dots, p_m)$

$$\begin{aligned} P(Y = j \mid \mathbf{X} = (x_1, x_2, \dots, x_m)) &= \frac{P(\mathbf{X} = (x_1, x_2, \dots, x_m) \mid Y = j) \cdot P(Y = j)}{P(\mathbf{X} = \mathbf{x})} \\ &= \frac{P(X_1 = x_1 \mid Y = j) \cdot \dots \cdot P(X_m = x_m \mid Y = j) \cdot P(Y = j)}{P(\mathbf{X} = \mathbf{x})}. \end{aligned}$$

$$\begin{aligned} \hat{y} &= \arg \max_{j \in \text{classes}} \frac{P(X_1 = x_1 \mid Y = j) \cdot \dots \cdot P(X_m = x_m \mid Y = j) \cdot P(Y = j)}{P(\mathbf{X} = \mathbf{x})} \\ &= \arg \max_{j \in \text{classes}} P(X_1 = x_1 \mid Y = j) \cdot \dots \cdot P(X_m = x_m \mid Y = j) \cdot P(Y = j). \end{aligned}$$

# Bayes Classifier: Naive One!

Let's dive into code!

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer

**Bayes Classifier**

Hyperparameters

Second approach:  
CNN

Sequence of vectors

How We Embed

Patterns with CNN

Dense Classifier

The End

# Hyperparameters

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Two important hyperparameters:

1. Size of the vocabulary;
2. Laplace/ Lidstone smoothing parameter( $\alpha$ ).

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier

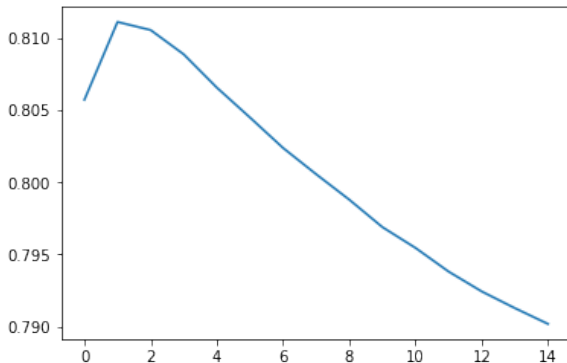
**Hyperparameters**

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

# Size of Vocabulary



It is convex! (to be completed)

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier

Hyperparameters

Second approach:  
CNN

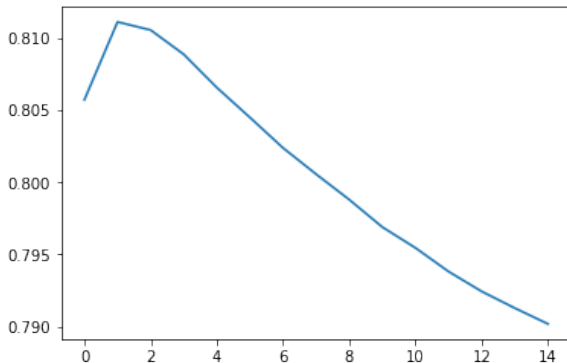
Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End

# Laplace/ Lidstone Smoothing Parameter( $\alpha$ )

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi



It is convex! (to be completed)

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier

Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End



# Grid Search

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features

No. of Classes

First approach:  
Naive Bayes

Count Vectorizer

Bayes Classifier

**Hyperparameters**

Second approach:  
CNN

Sequence of vectors

How We Embed

Patterns with CNN

Dense Classifier

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features

No. of Classes

First approach:  
Naive Bayes

Count Vectorizer

Bayes Classifier

Hyperparameters

Second approach:  
CNN

Sequence of vectors

How We Embed

Patterns with CNN

Dense Classifier

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features

No. of Classes

First approach:  
Naive Bayes

Count Vectorizer

Bayes Classifier

Hyperparameters

Second approach:  
CNN

Sequence of vectors

**How We Embed**

Patterns with CNN

Dense Classifier

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
**Patterns with CNN**  
Dense Classifier

The End

# Temp Frame

This slide is temp.

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features

No. of Classes

First approach:  
Naive Bayes

Count Vectorizer

Bayes Classifier

Hyperparameters

Second approach:  
CNN

Sequence of vectors

How We Embed

Patterns with CNN

**Dense Classifier**

The End

# Thanks for your attention!

Codes in slides (in my GitHub):(github link)

Divar posts dataset:(divar link)

Any questions?

An Experience  
with Text  
Classification in  
Datadays 2019

Majid Hajiheidari,  
Amirmohammad  
Asadi

Introduction:  
Divar Dataset

The Problem:  
Categorization

Features  
No. of Classes

First approach:  
Naive Bayes

Count Vectorizer  
Bayes Classifier  
Hyperparameters

Second approach:  
CNN

Sequence of vectors  
How We Embed  
Patterns with CNN  
Dense Classifier

The End