«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

ИТОГОВАЯ АТТЕСТАЦИОННАЯ РАБОТА

Прогнозирование и анализ оттока пользователей по сообщениям в соцсети

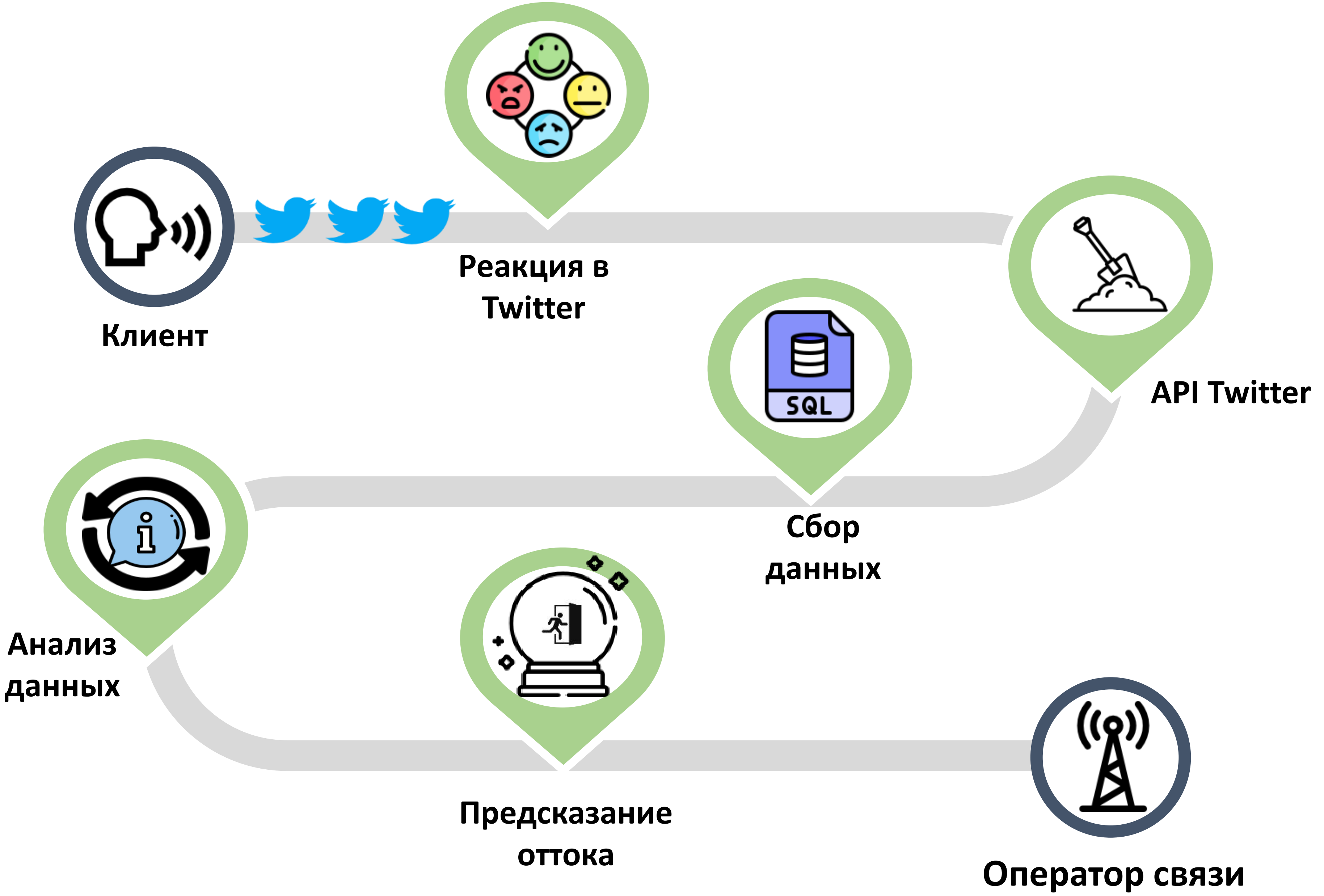
Выполнил:

Крылов Сергей

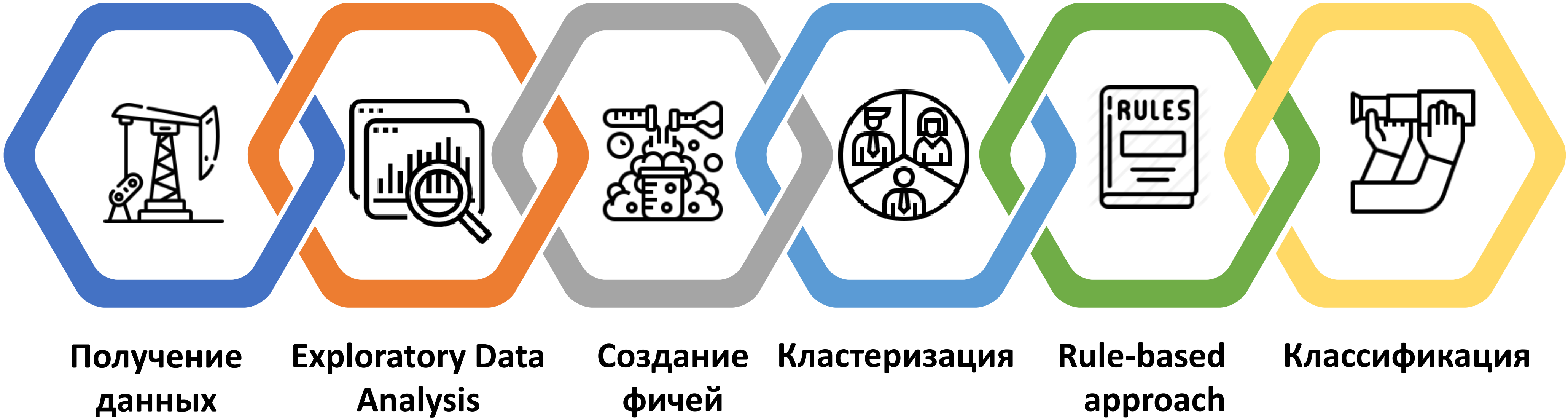
Руководитель:

Селезнев Артем

Актуальность проблемы



Алгоритм работы



Получение данных

Exploratory Data Analysis

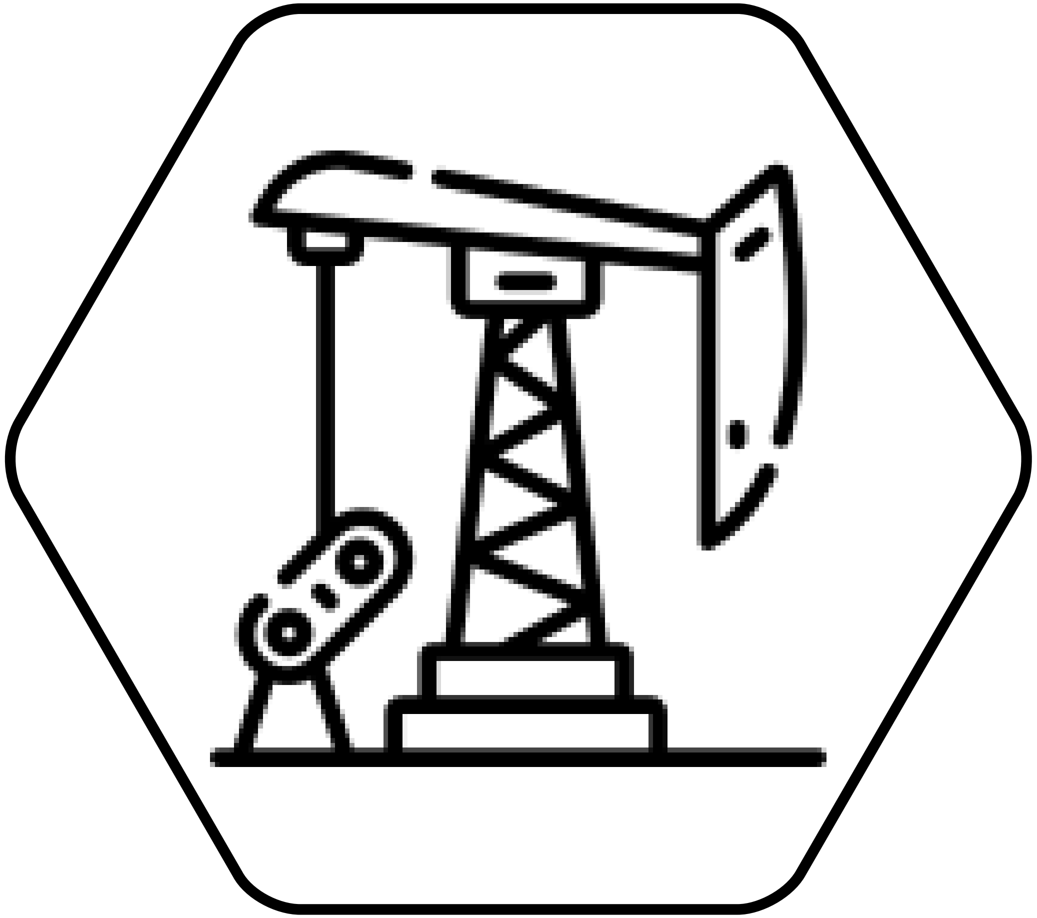
Создание фичей

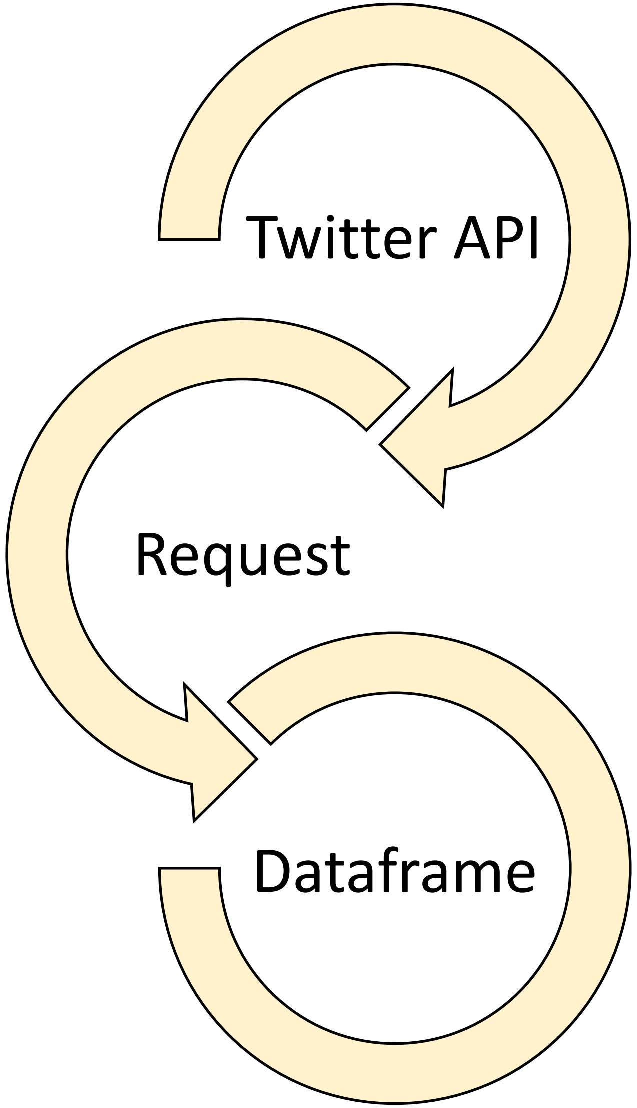
Кластеризация

Rule-based approach

Классификация

Получение данных





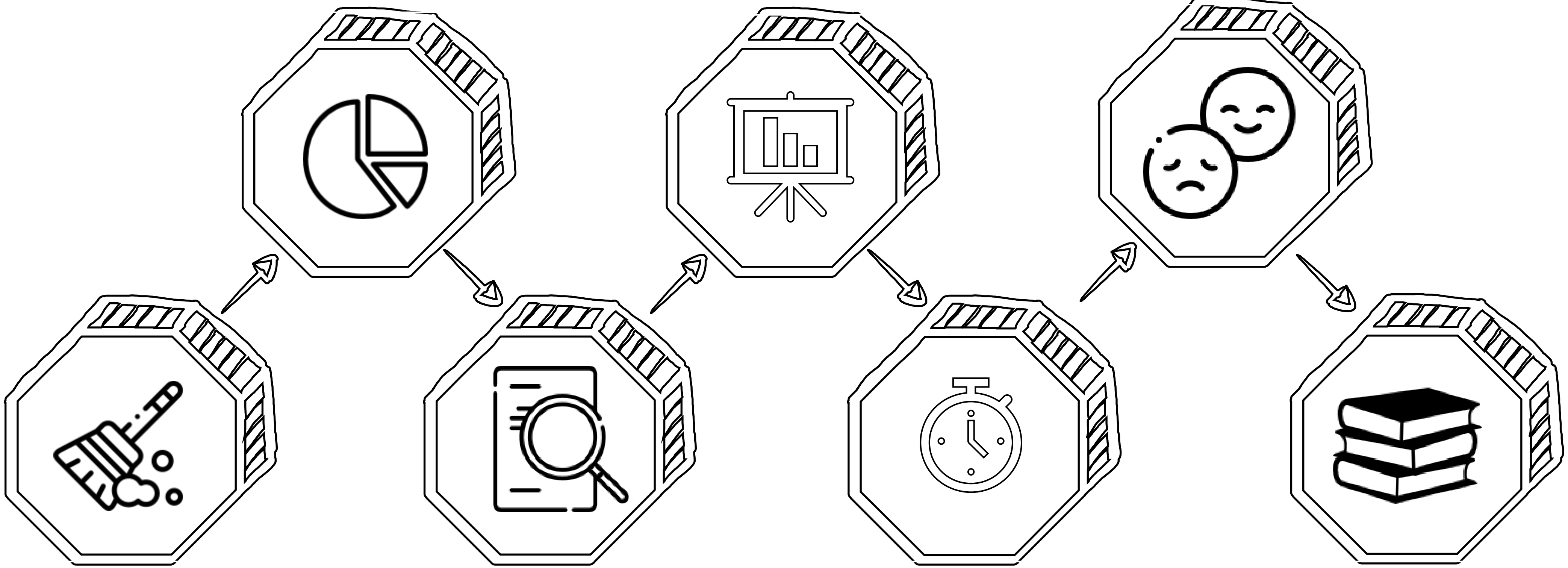
```
tweepy.OAuthHandler(tokens)
set_access_token
tweepy.API(auth)
```

```
tweepy.Cursor(api.search, q=searchString, lang='en’)
```

	screen_name	date_time	location	text
0	TheSkubis	2021-03-30 23:59:19	Pennsylvania, USA	@VerizonSupport I have issues with closed capt...
1	VerizonSupport	2021-03-30 23:56:55		@_carolinek This could be due to regional rest...
2	VerizonSupport	2021-03-30 23:55:30		@tvmurray We'll be happy to help with anything...

Exploratory Data Analysis

Exploratory Data Analysis



preprocessing

сегменты

df.describe

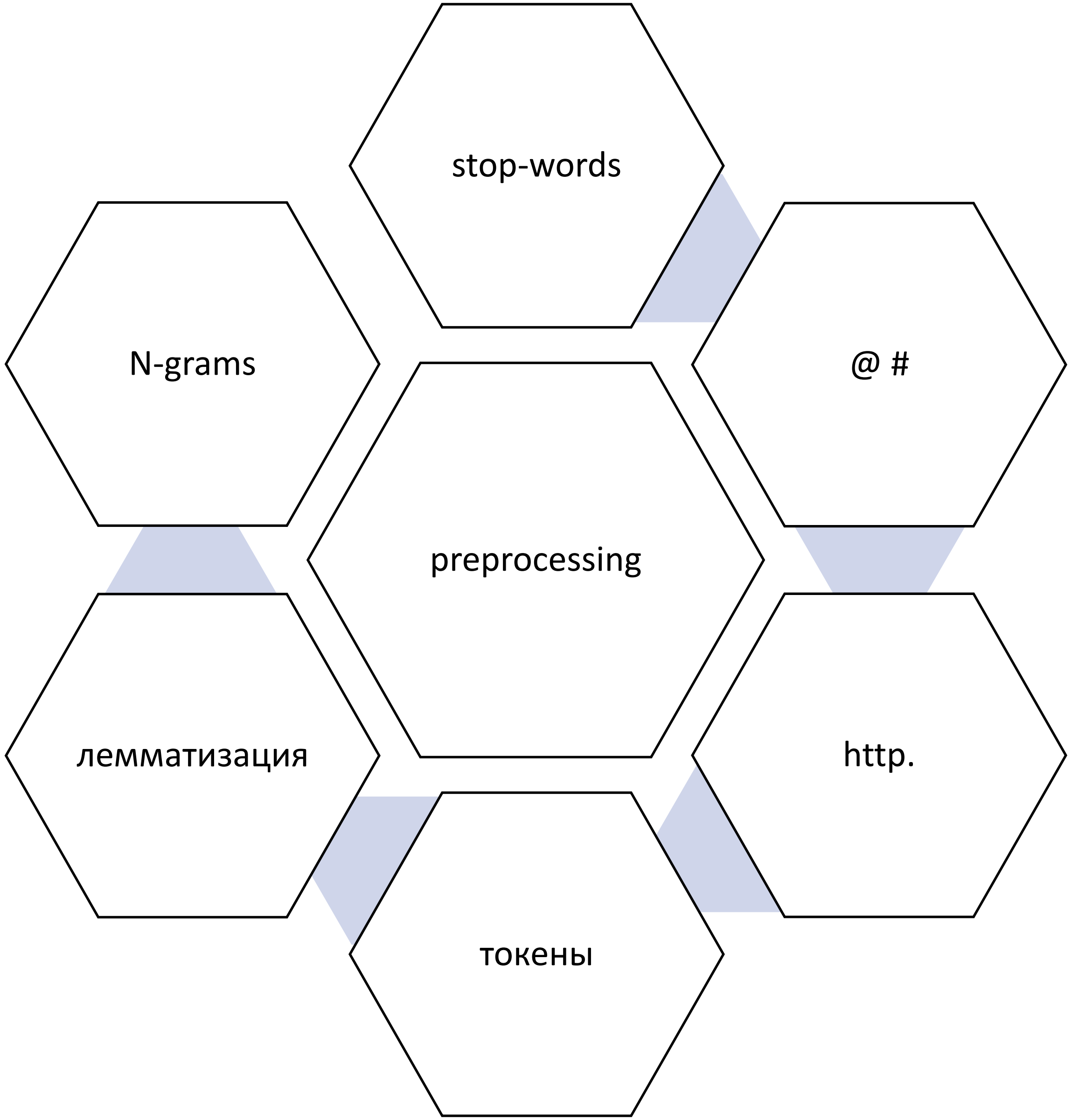
частоты

время

сентименты

топики

Preprocessing



preprocessing

stop-words

@ #

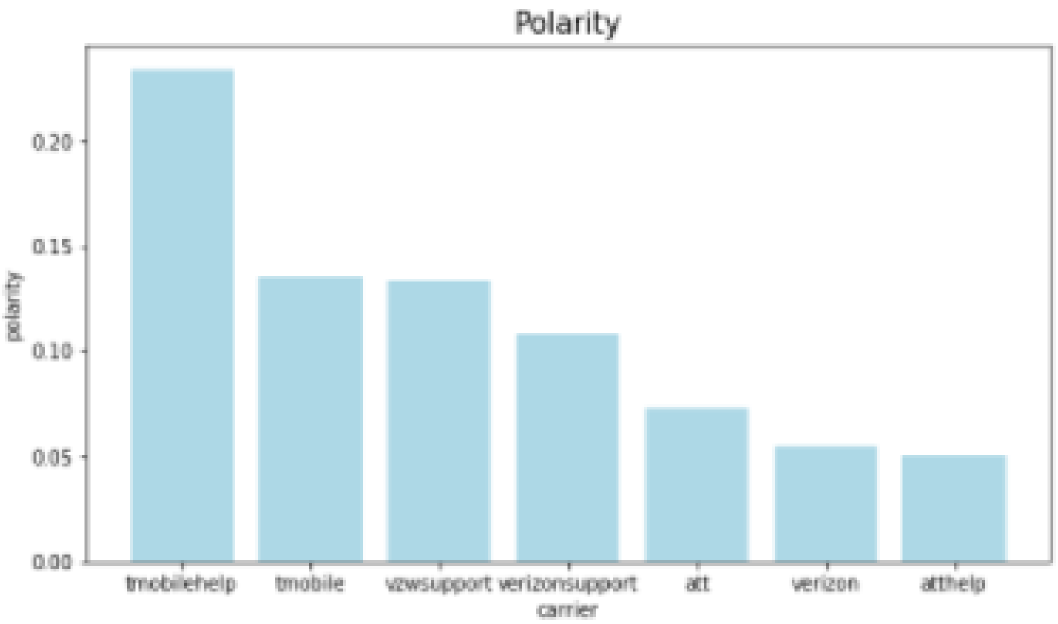
http.

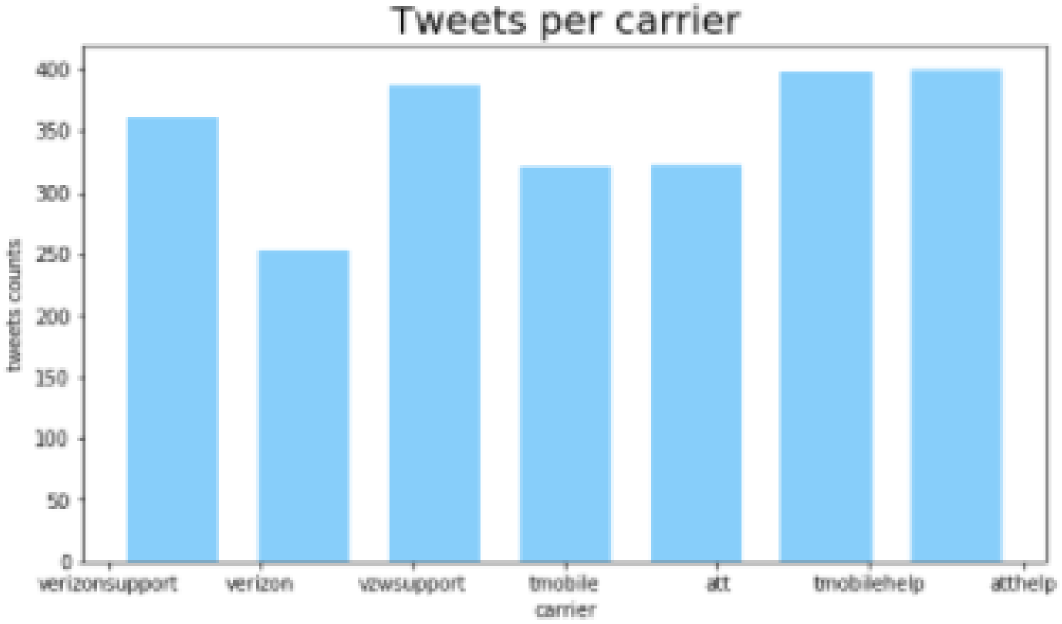
токены

лемматизация

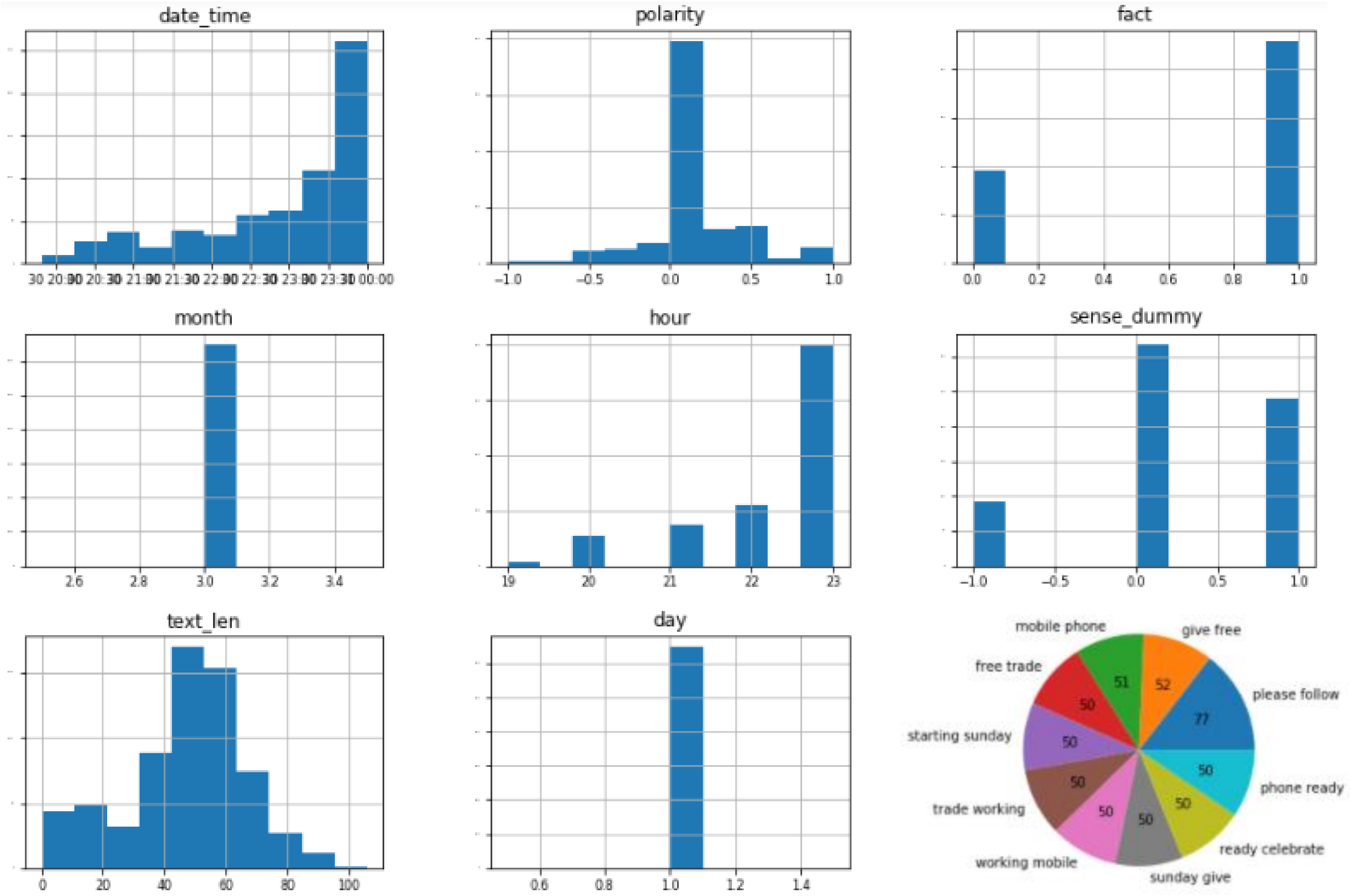
N-grams

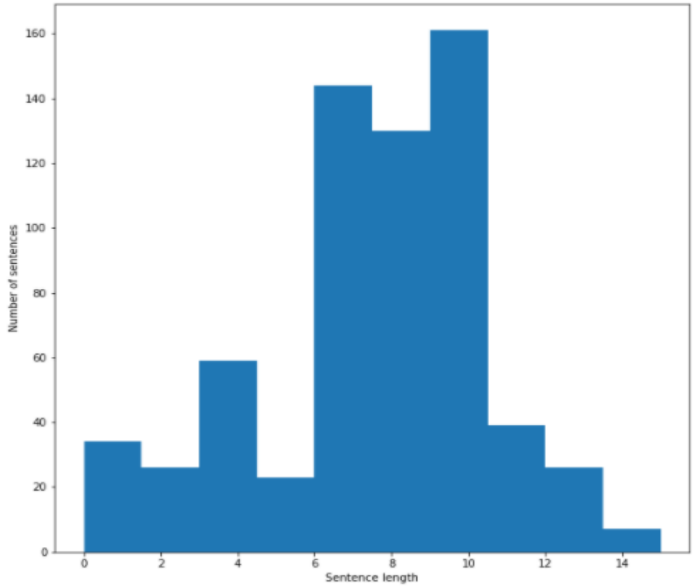
Сегменты

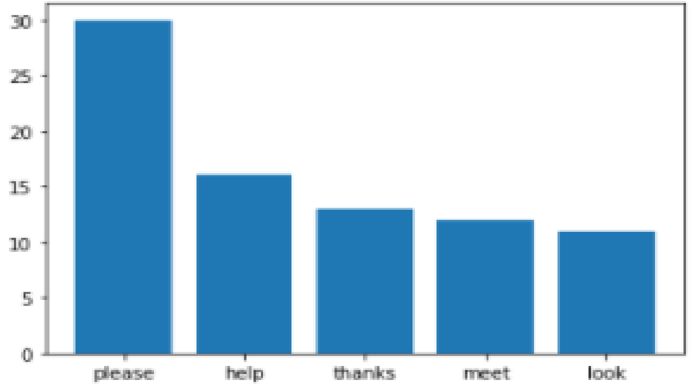

Polarity
polarity
0.00
0.05
0.10
0.15
0.20
tmobilehelp
tmobile
vzwsupport
verizonsupport
att
verizon
atthelp
carrier


Tweets per carrier
tweets counts
0
50
100
150
200
250
300
350
400
verizonsupport
verizon
vzwsupport
tmobile
att
tmobilehelp
atthelp
carrier

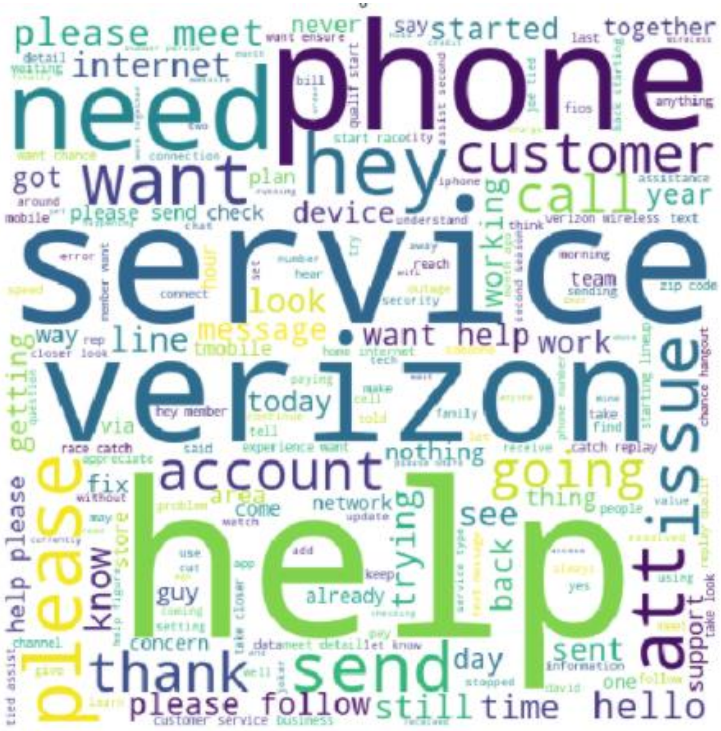
df.describe


date_time
30 20:30
30 20:30
30 21:00
30 21:30
30 22:00
30 22:30
30 23:00
30 23:30
31 00:00
polarity
−1.0
−0.5
0.0
0.5
1.0
fact
0.0
0.2
0.4
0.6
0.8
1.0
month
2.6
2.8
3.0
3.2
3.4
hour
19
20
21
22
23
sense_dummy
−1.0
−0.5
0.0
0.5
1.0
text_len
0
20
40
60
80
100
day
0.6
0.8
1.0
1.2
1.4
mobile phone
give free
free trade
please follow
starting sunday
phone ready
trade working
ready celebrate
working mobile
sunday give
51
52
50
77
50
50
50
50
50
50

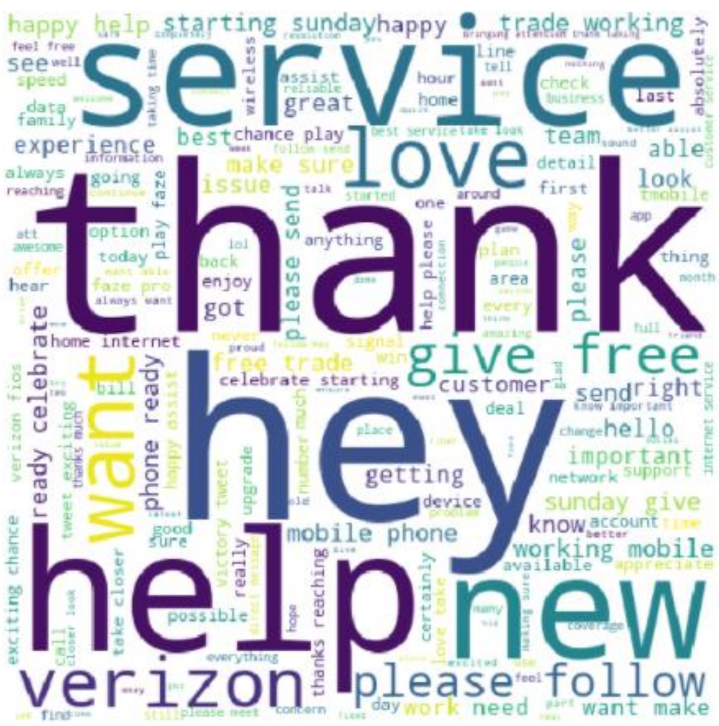

Number of sentences
0
20
40
60
80
100
120
140
160
Sentence length
0
2
4
6
8
10
12
14


0
5
10
15
20
25
30
please
help
thanks
meet
look

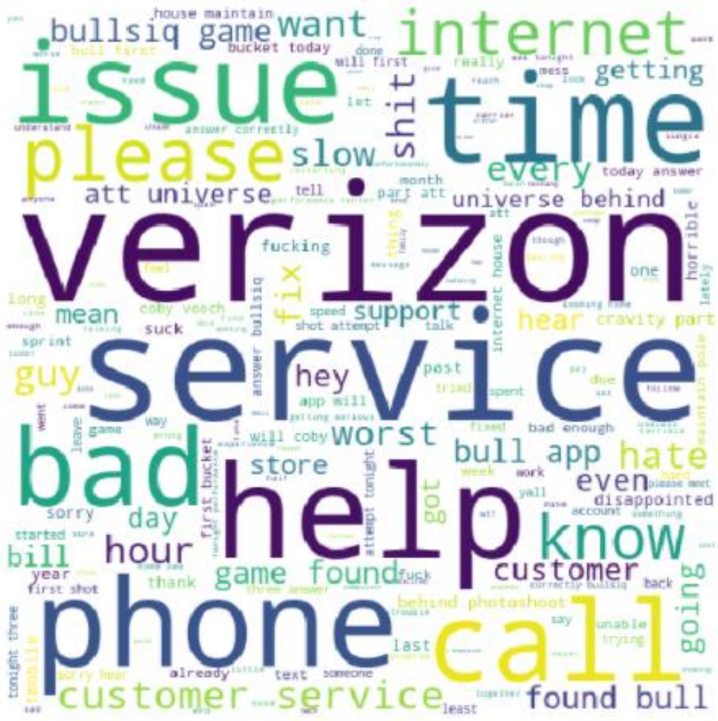
Анализ частот слов по сентиментам



😐



🙂



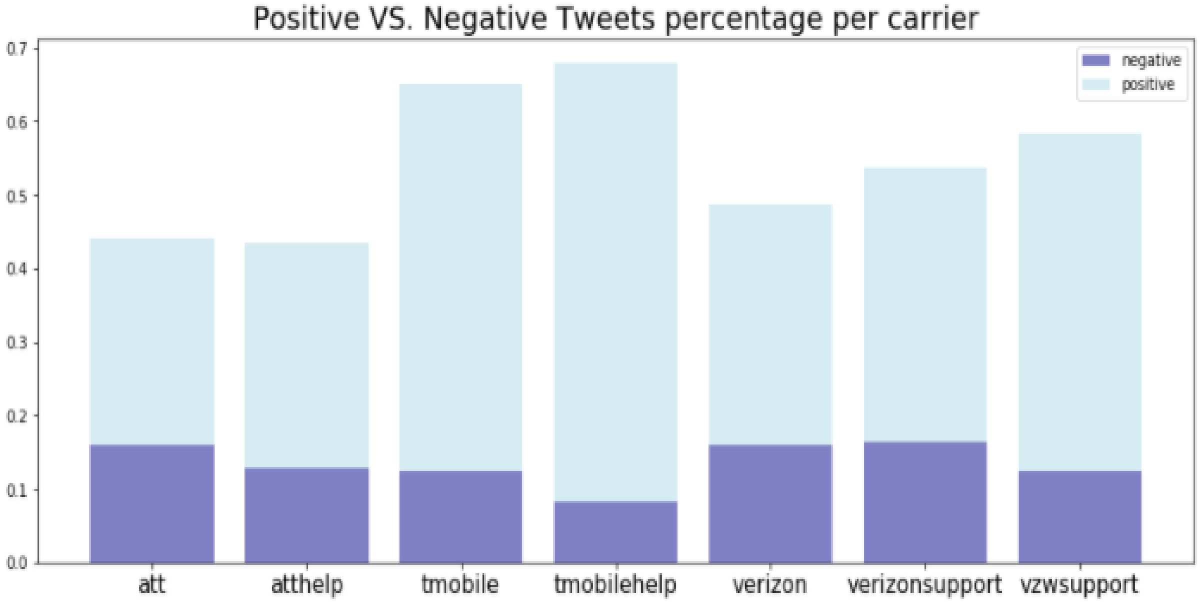
☹

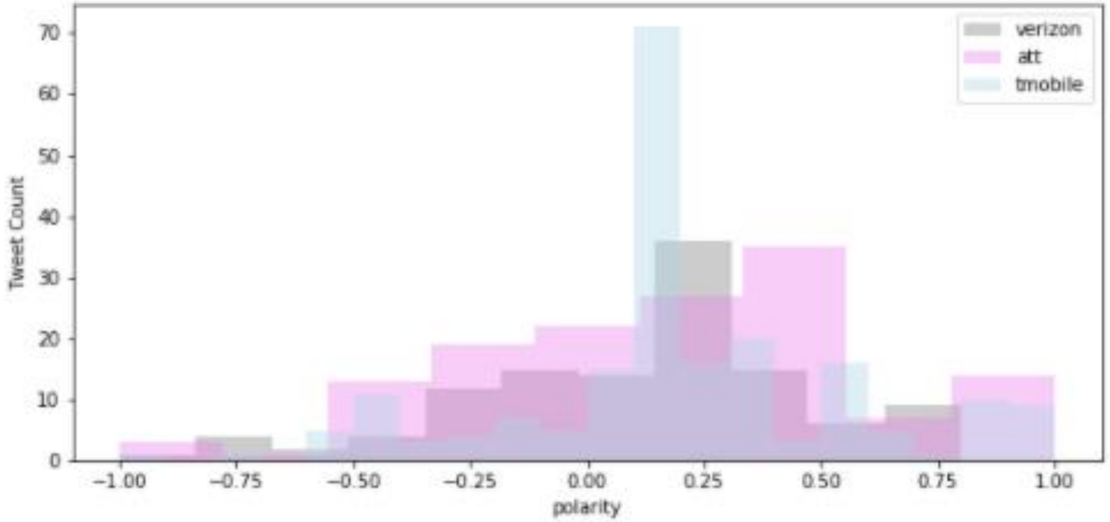
Время

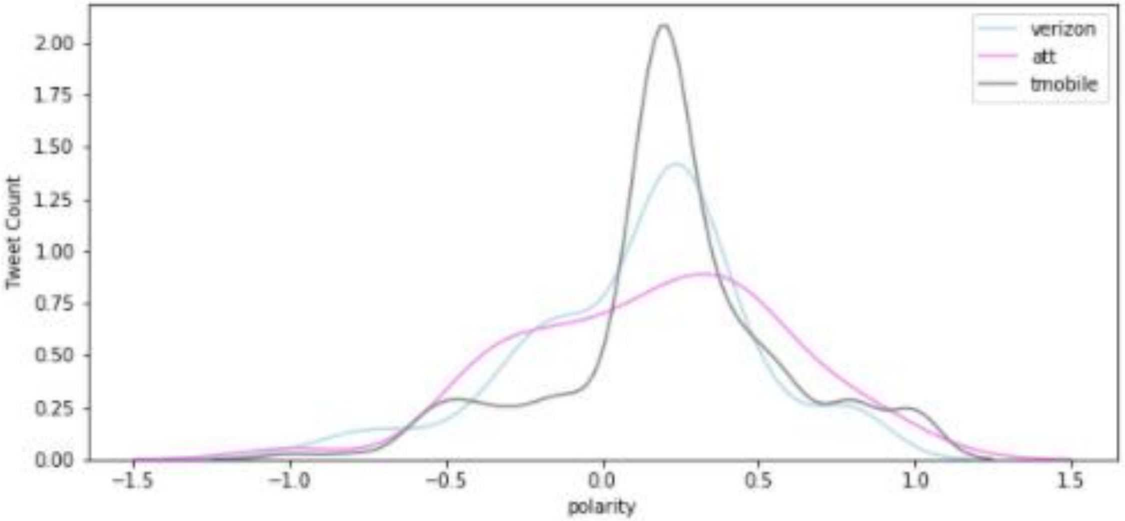
- НУЖНО БОЛЬШЕ ДАННЫХ

ACTUAL

Сентименты


Positive VS. Negative Tweets percentage per carrier
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
negative
positive
att
atthelp
tmobile
tmobilehelp
verizon
verizonsupport
vzwsupport


verizon
att
tmobile
Tweet Count
70
60
50
40
30
20
10
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
polarity


verizon
att
tmobile
Tweet Count
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
polarity

Топики

1 help get got keep team verizon full per pas attempt

2 help cut time great get funny number given experiencing thanks

3 wireless verizon need best direct way get feel please work

4 using customer account hear app true tvision try need att

5 sent thanks told elc hey would keep needed tried order

6 help send month want take look code zip detail please

7 need give service classroom get hello please type work today

8 account month vacation billed suspension verizon att reaching call since

9 customer internet please thanks service working local victim tornado provider

10 help please happy send jersey follow detail today assist look

11 check device get new hour call detail data pay wireless

12 want make sure experience service need help fuck getting sorry

13 phone know verizon around hanging minute anyone spare internet let

14 verizon thank great live fios lost service hear att home

15 would love please look team meet back help get follow

16 tweet prank every tap year donating time verizon customer million

17 please verizon hello could one get send customer issue use

18 guy day supposed verizon back say long sprint past ordered

19 help always family happy service love card ever fios still

20 auto get cut said like thank maybe hold paying hell

Создание фичей

время

Применение sin/cos трансформации

TF-IDF

Отбор топ-80%

RFM-analysis

Проекция RFM на Recency, frequency, polarity

location_proba

Вероятность города

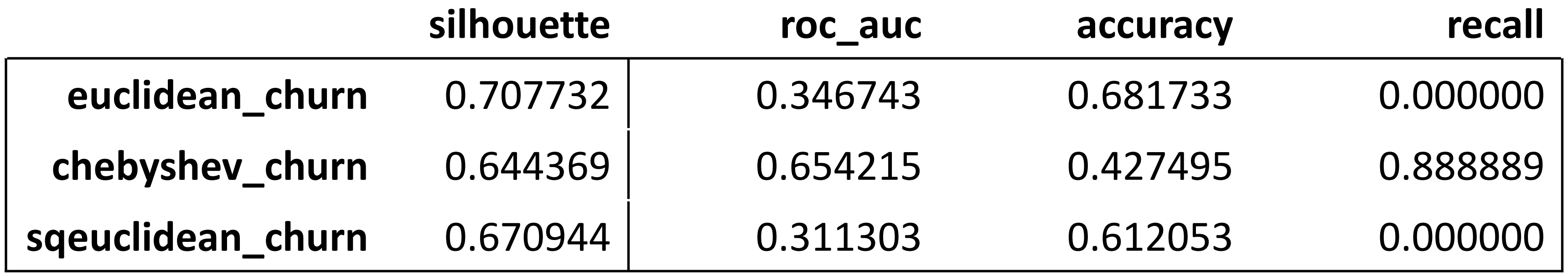
fact

Textblob. subjectivity

topic

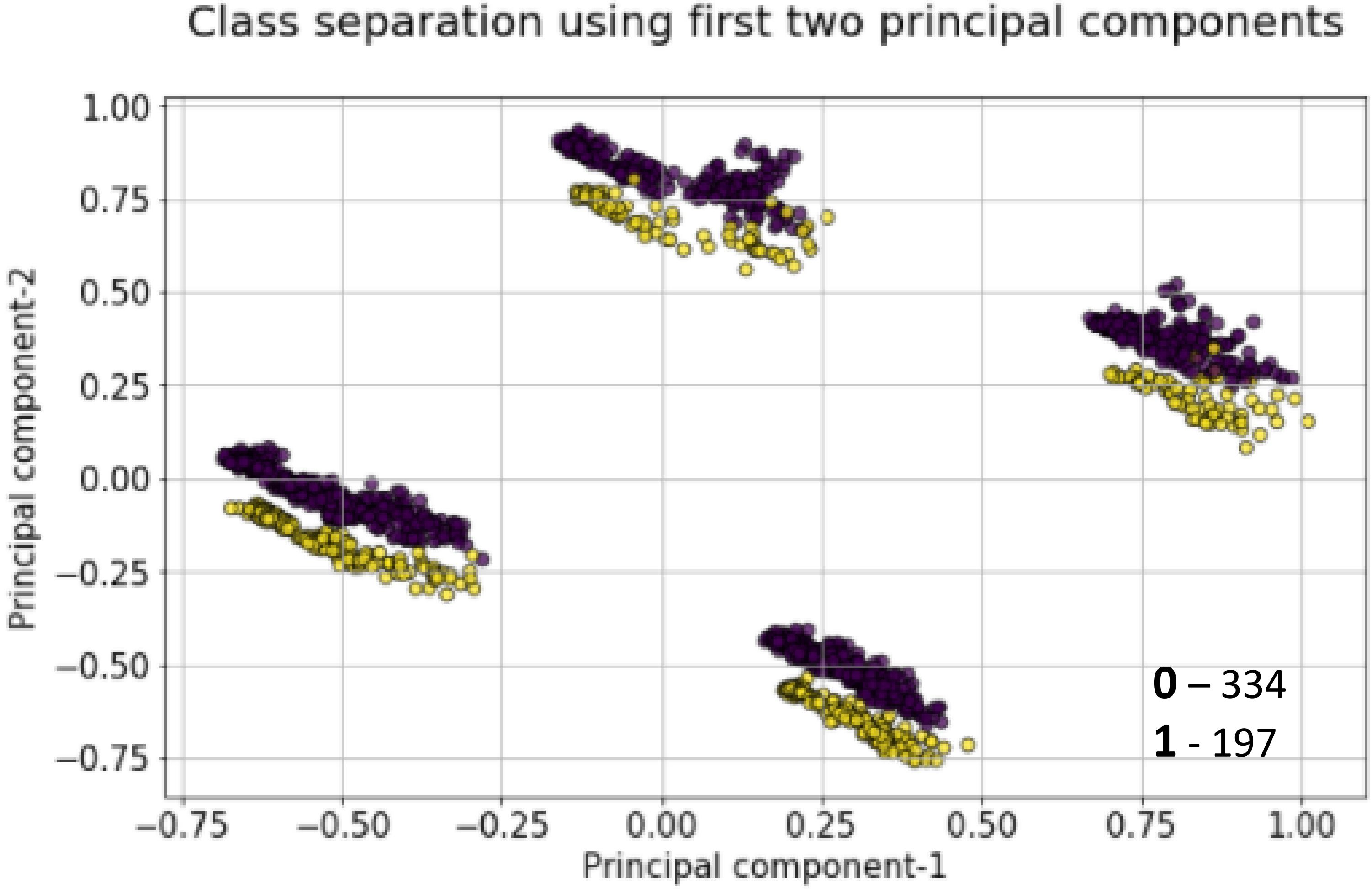
Номер топика из topic_modeling

Кластеризация



	silhouette	roc_auc	accuracy	recall
euclidean_churn	0.707732	0.346743	0.681733	0.000000
chebyshev_churn	0.644369	0.654215	0.427495	0.888889
sqeuclidean_churn	0.670944	0.311303	0.612053	0.000000

ACTUAL

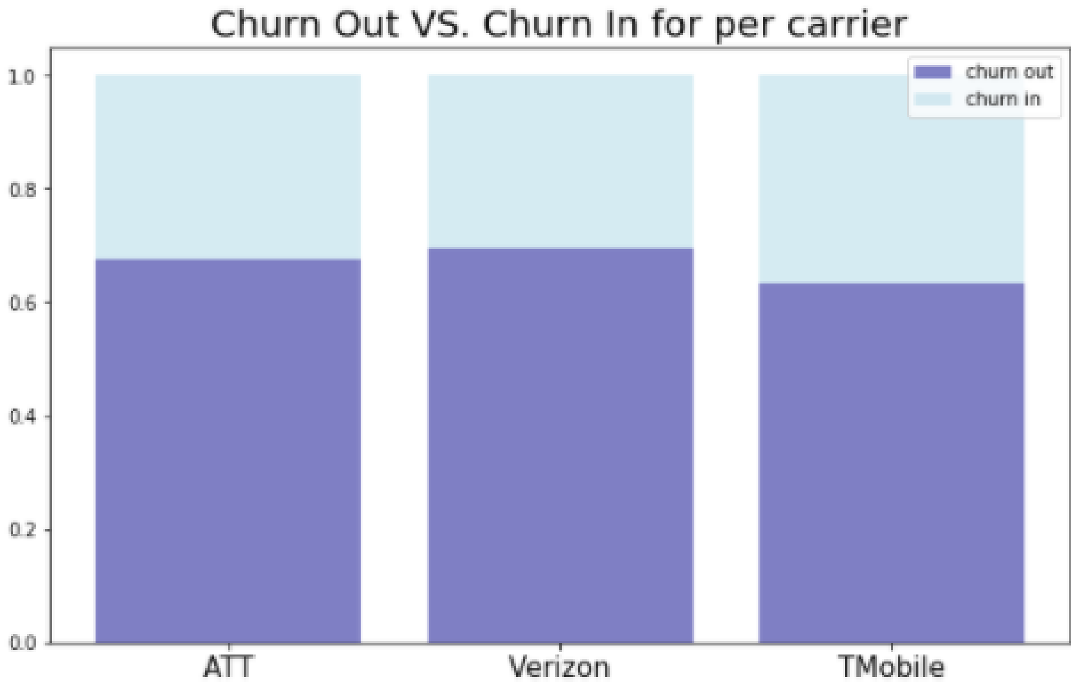

Class separation using first two principal components
Principal component-2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
Principal component-1
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

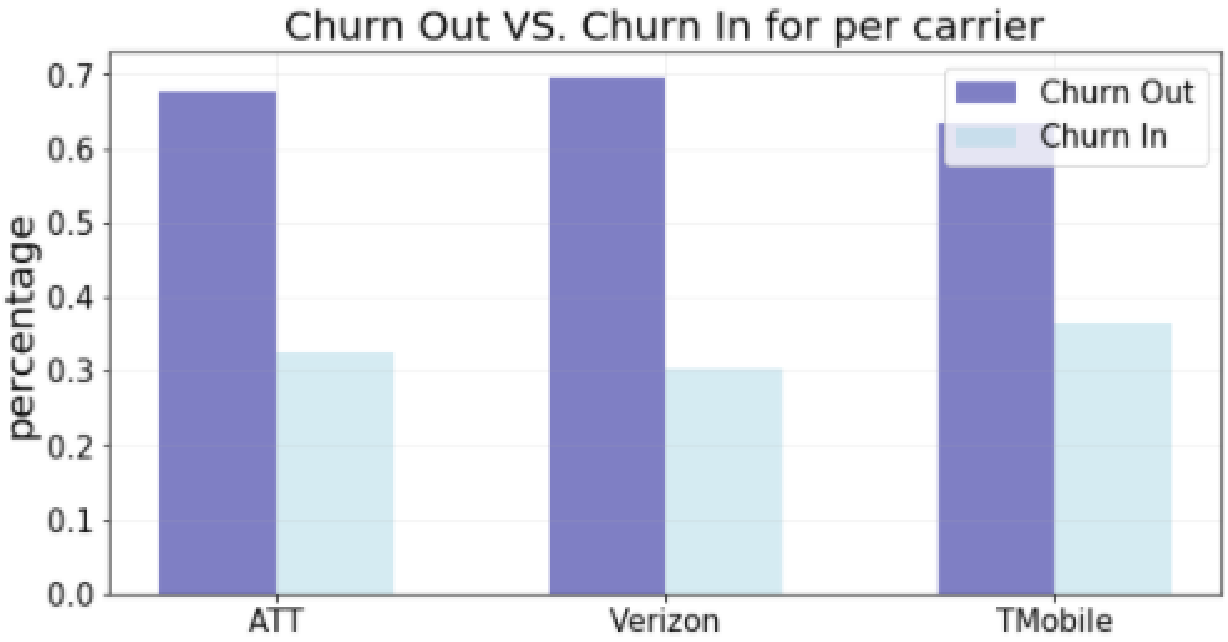
0 – 334

1 - 197

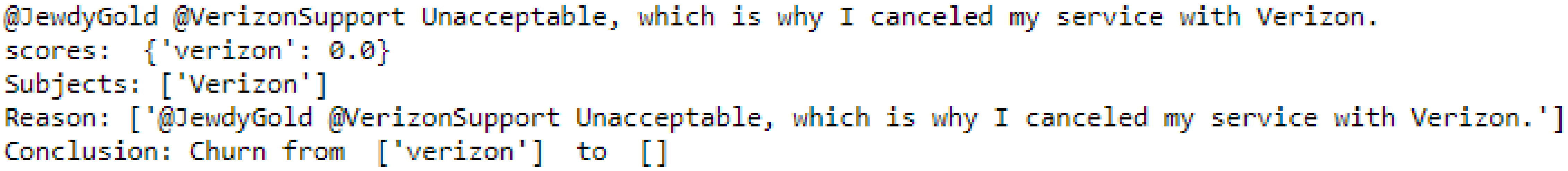
Ruled-based approach

Rule		**churny words**	**churny reasons**
‘from’	-0.8	switch , transfer	service, better, experts, disney
‘to’	+0.8	free, come	network, wifi, price, tower, coverage
‘with’	+0.5	come, change	family, horrible, awfull, slow, free of charge
‘like’	+0.5	leave, stay	worst, worse, bad, price, money
‘Disney’ & ‘Verizon’	+0.5	welcome, goodbye	can\'t stand, promise
sentiment	Count(carrier) x polarity		
score < 0	Отток_от.append(оператор)		
Score > 0	Отток_к.append(оператор)		


Churn Out VS. Churn In for per carrier
churn out
churn in
1.0
0.8
0.6
0.4
0.2
0.0
ATT
Verizon
TMobile

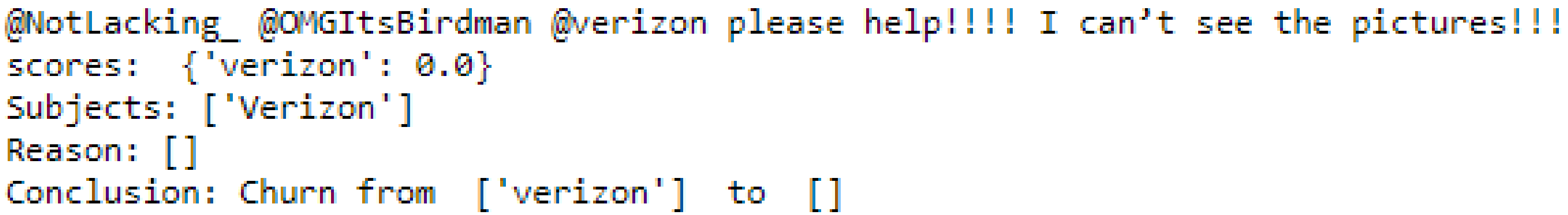

Churn Out VS. Churn In for per carrier
Churn Out
Churn In
percentage
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
ATT
Verizon
TMobile

Хороший пример



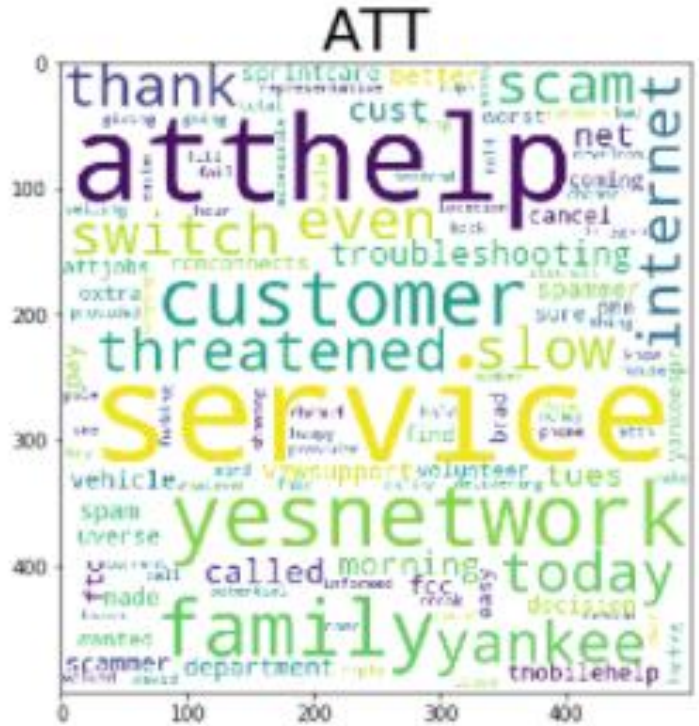
```
@JewdyGold @VerizonSupport Unacceptable, which is why I canceled my service with Verizon.
scores:  {'verizon': 0.0}
Subjects: ['Verizon']
Reason: ['@JewdyGold @VerizonSupport Unacceptable, which is why I canceled my service with Verizon.']
Conclusion: Churn from  ['verizon']  to  []
```

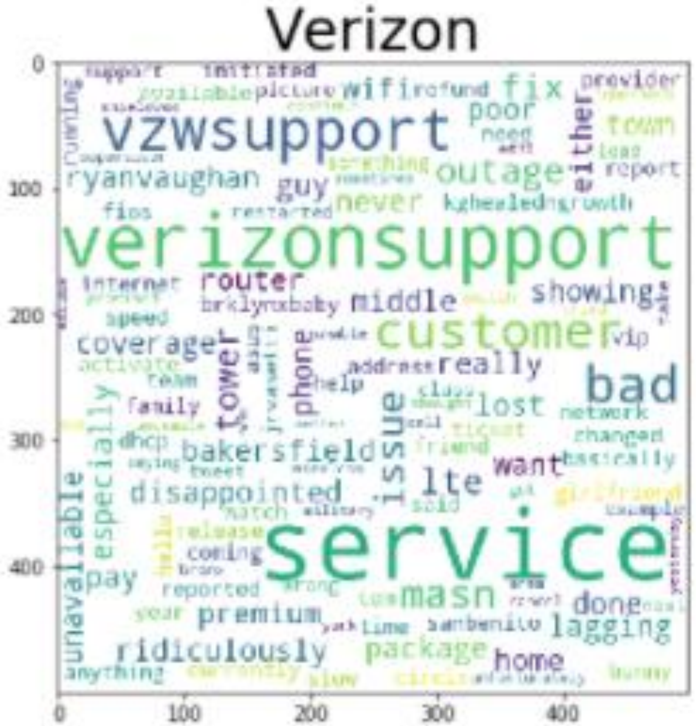
Пример не очень

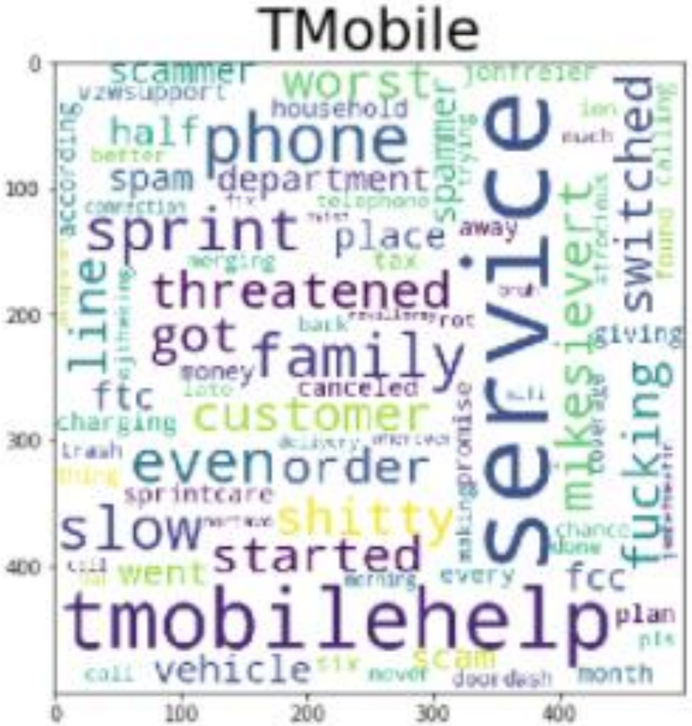


```
@NotLacking_ @OMGItsBirdman @verizon please help!!!! I can’t see the pictures!!!
scores:  {'verizon': 0.0}
Subjects: ['Verizon']
Reason: []
Conclusion: Churn from  ['verizon']  to  []
```

Облако слов причин оттока


ATT

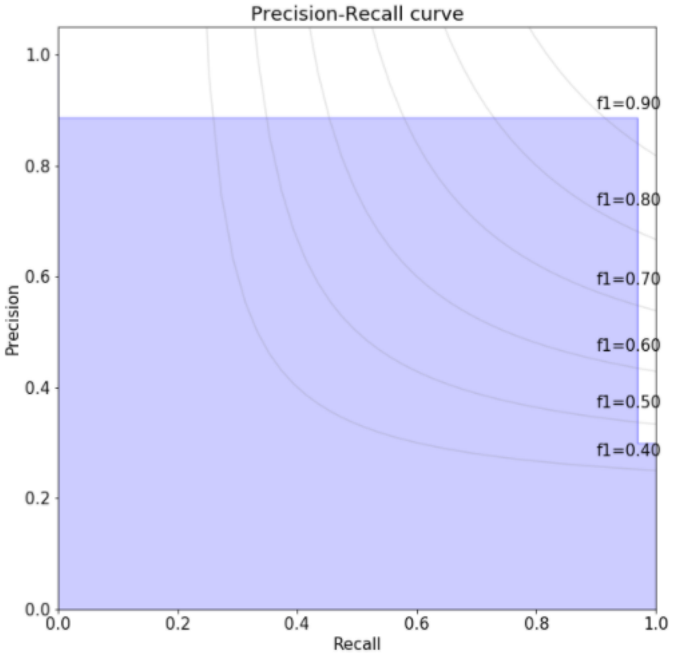

Verizon

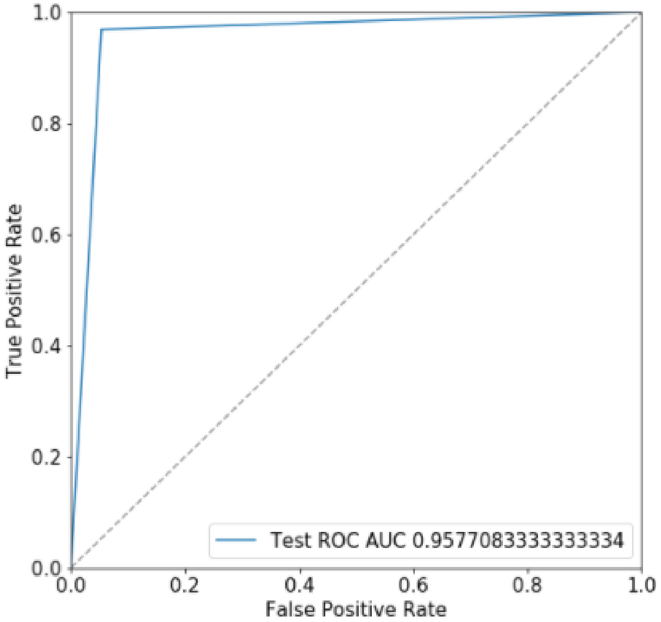

TMobile

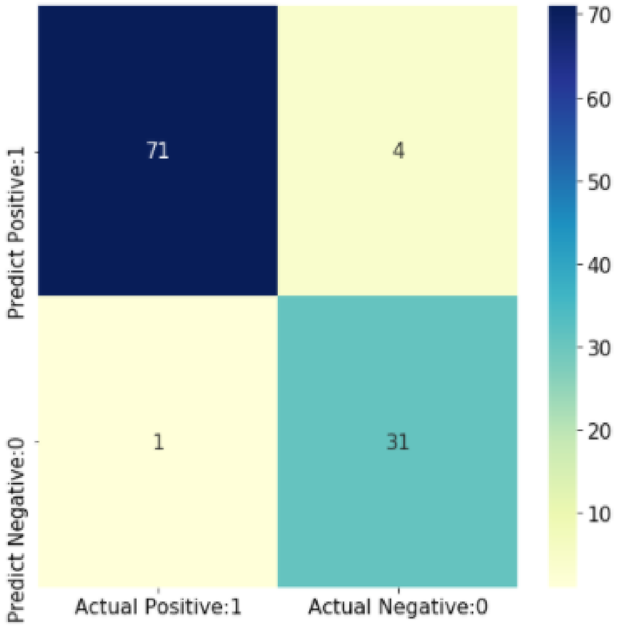
Классификация

model	roc_auc
Naïve Bayes	0.546875
Logistic regression	0.750625
LogRegression + grid_search	0.797500
LogRegression + MinMaxScaler	0.792917
LogRegression + grid_search + MinMaxScaler	0.817500
Decision Tree	0.815208
LightGBM	0.857708

LightGBM







	precision	recall	score	support
0	0.99	0.95	0.97	75
1	0.89	0.97	0.93	32
accuracy			0.95	107
macro avg	0.94	0.96	0.95	107
weighted avg	0.96	0.95	0.95	107

	train	test
Cross_validate	1	0.99

спасибо за внимание!