Baseline

(1) автоматическое деление в CountVectorizer и TfidfVectorizer, убираются ссылки
(2) стоп-слова при l1 penalty зануляются, если они не важны, в обратном случае, нам они важны
(2) нет
(3) CountVectorizer и TfidfVectorizer
(4) LogisticRegression(penalty='l1')

Улучшения

(1) разбиваем на токены **nltk.tokenize.TweetTokenizer** с хорошим учетом знаков препинания, которые выражают тональность
(2) стопслова могут быть значимыми
(2) pymorpy2 или стеммер (SnowballStemmer из nltk)
(3) TF-IDF с разными параметрами
(4) разные классификаторы

1. Токенизатор

В твитах особенно важно учитывать знаки препинания, что важно при оценке тональности. В модуле nltk есть специальный адаптированный токенизатор.

2. Лемматизация

Два варианта:

- лемматизация
  лемматизация с помощью pymorphy, так как он обучался на более современных текстах, хорошо работает с нестандартными формами (которые часты в онлайн-речи), работает быстрее (по опыту).
- стемминг
  SnowballStemmer(‘russian’) из nltk

3. TF-IDF

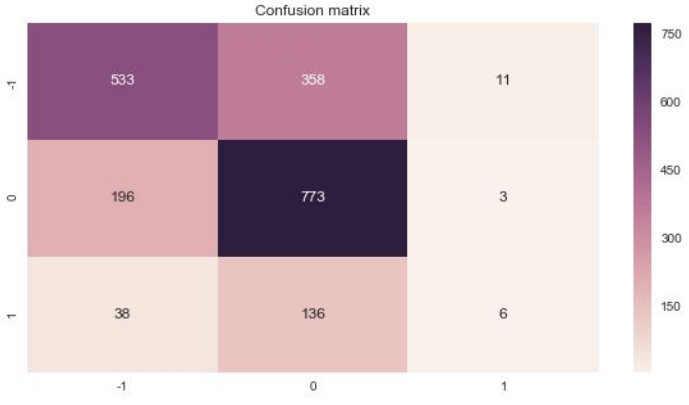
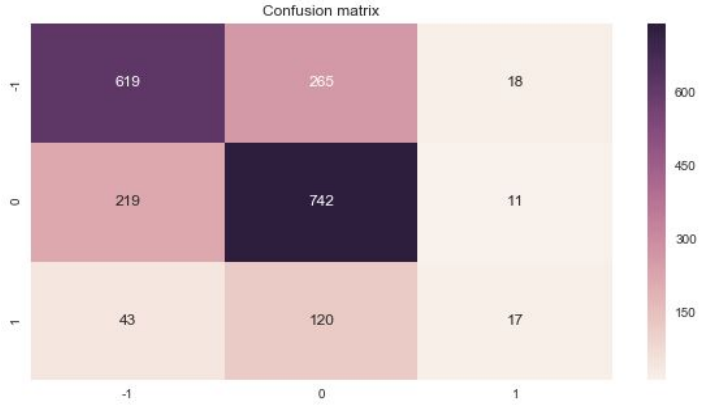
-

4. Классификаторы

- SVC(class_weight='balanced',kernel = 'rbf')
  GridSearchCV с f1_weighted оценкой
- SVC(class_weight='balanced',kernel = 'rbf',decision_function_shape='ovo')
  GridSearchCV с той же оценкой
- RandomForestClassifier(n_estimators=50, random_state=23,
  class_weight='balanced')
- LogisticRegressionCV(Cs=list(np.power(10.0, np.arange(-20, 20))),
  scoring='f1_weighted',
  class_weight='balanced',
  multi_class='multinomial',
  random_state=23)

5. Подбор параметров
- подбор параметров с помощью GridSearchCV
- В LogisticRegressionCV задающий параметр Cs=list(np.power(10.0, np.arange(-20, 20)))
6. Отбор параметров
- отбор параметров с помощью логистической регрессии с L1 для минимизации нерелевантных признаков и их сокращения, что важно для GridSearchCV, где хотелось бы работать быстрее (а не на тысячах признаков)

Результаты baseline

CountVectorizer	TfidfVectorizer
precision recall f1-score support -1 0.69 0.59 0.64 902 0 0.61 0.80 0.69 972 1 0.30 0.03 0.06 180 avg / total 0.62 0.64 0.61 2054 Макросредняя F1 мера - 0.46306421211286786 Микросредняя F1 мера - 0.6387536514118792 	precision recall f1-score support -1 0.70 0.69 0.69 902 0 0.66 0.76 0.71 972 1 0.37 0.09 0.15 180 avg / total 0.65 0.67 0.65 2054 Макросредняя F1 мера - 0.5172604008633273 Микросредняя F1 мера - 0.6708860759493671 

Лучший вариант

Предобрботка

TweetTokenizer
SnowballStemmer
TfidfVectorizer(ngram_range =(1,2), token_pattern='\S+')

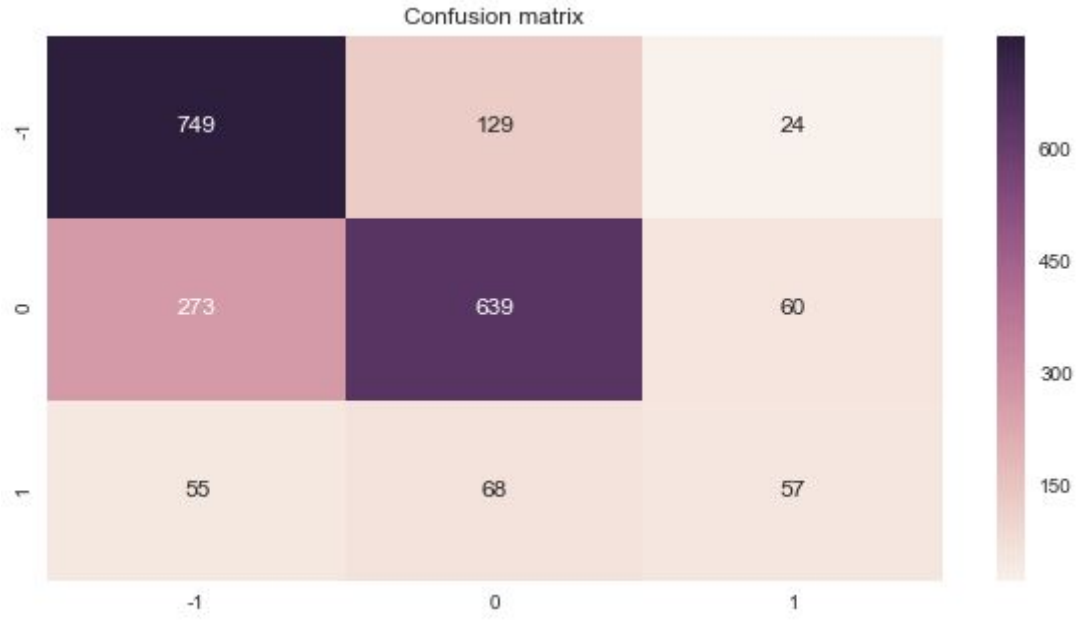
	precision	recall	f1-score	support
-1	0.70	0.83	0.76	902
0	0.76	0.66	0.71	972
1	0.40	0.32	0.36	180
avg / total	0.70	0.70	0.70	2054

Макросредняя F1 мера - 0.6063155158358028
Микросредняя F1 мера - 0.7035053554040895

Для сравнения:

CV
Макр F1 мера - 0.4631
Микро F1 мера - 0.6388

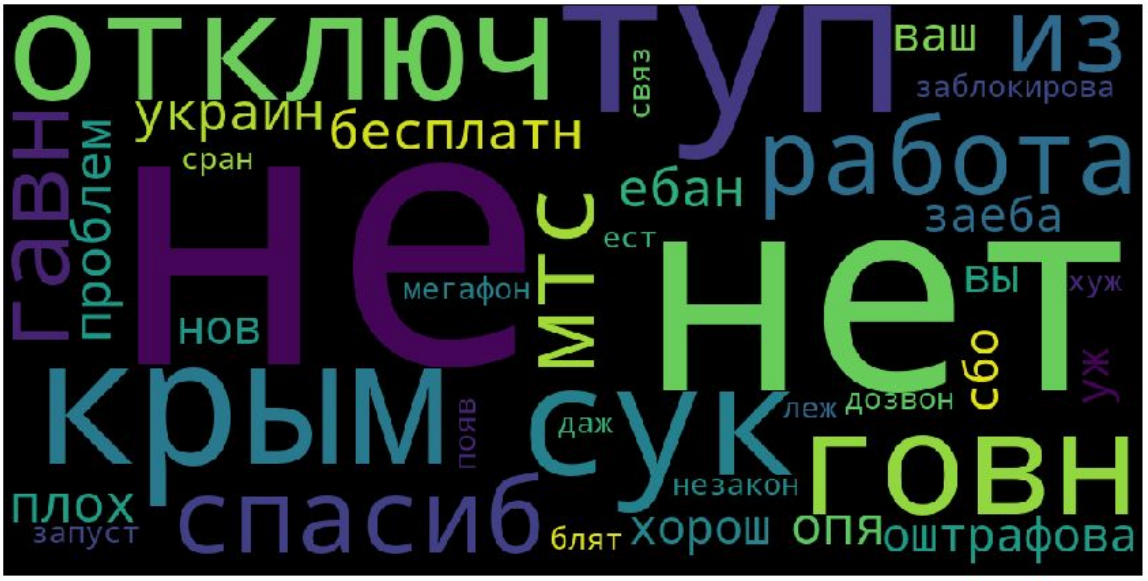
TFIDF
Макро F1 мера - 0.5173
Микро F1 мера - 0.6709



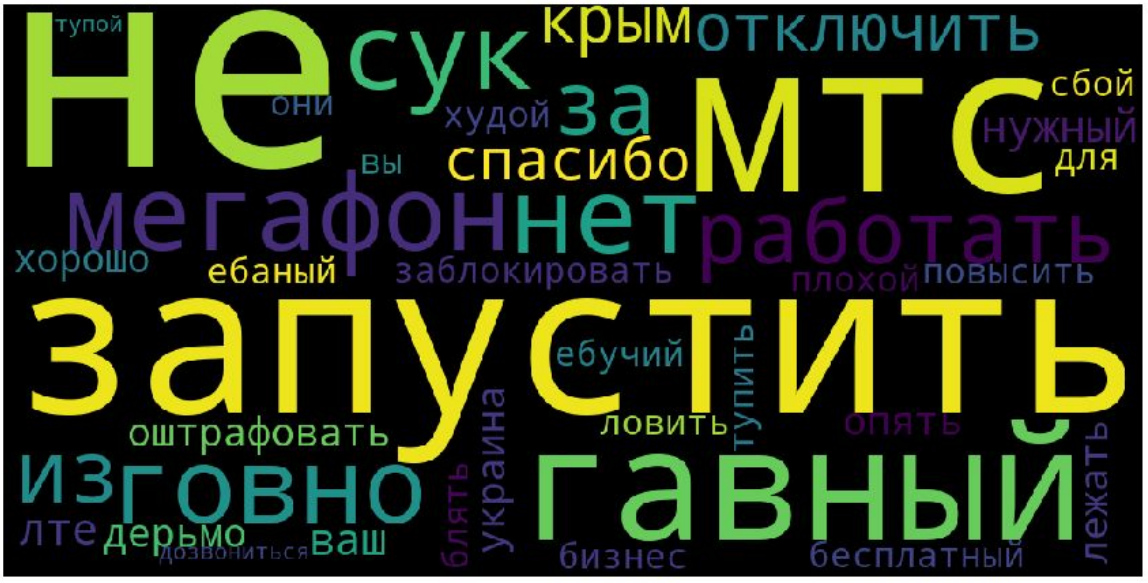
Лучше выделяется отрицательный класс, увеличился TP для 1 класса.

Релевантные признаки

Стемминг



Лемматизация



Признаки из лучшей модели

-1		0		1	
не	10.099923	. #билайн	7.040962	**спасиб**	8.283071

нет	7.629972	beeline_rus	7.014003	**любл**	7.293637
туп	6.796865	#новост сам	5.080018	узбекиста	6.706562
крым	6.524265	ru	4.971660	связ #новост	6.694028
сук	6.419232	доллар	4.911762	**хорош**	6.614154
отключ	6.388011	прос	4.714097	**бесплатн**	6.173741
говн	6.342778	. ru	4.658166	**защит**	5.992362
.	6.229170	_beeline_kz	4.647269	**доступн**	5.932602
не работа	6.132987	: билайн	4.574640	**мегафон запуст**	5.742683
из-з	5.716777	карт	4.548090	**запуст**	5.718056
гавн	5.653767	. ^	4.542625	lte	5.665452
мтс-украин	5.554610	**инструкц**	4.456220	**:)**	5.635844
опя	5.430235	^	4.448988	**тепер**	5.425020
плох	5.403415	**настройк**	4.417643	**лучш**	5.402298
ебан	5.268444	клиент #читаювзаимн	4.408769	**заработа**	5.397983
проблем	5.263139	для	4.380005	мил	5.241363
оштрафова	5.113329	на трет	4.290498	**4g**	5.010560
заеба	4.946864	телефон	4.260406	**расширя**	4.993089
сбо	4.923528	: вчер	4.212640	**: в**	4.931954
уж	4.876214	спасиб .	4.194919	**появ**	4.869797

Для нейтрального класса не очень понятно, почему эти ключевые слова. Они коллекционные (актуальны в рамках этого набора данных). Для отрицательного и положительного довольно интуитивно понятные слова.