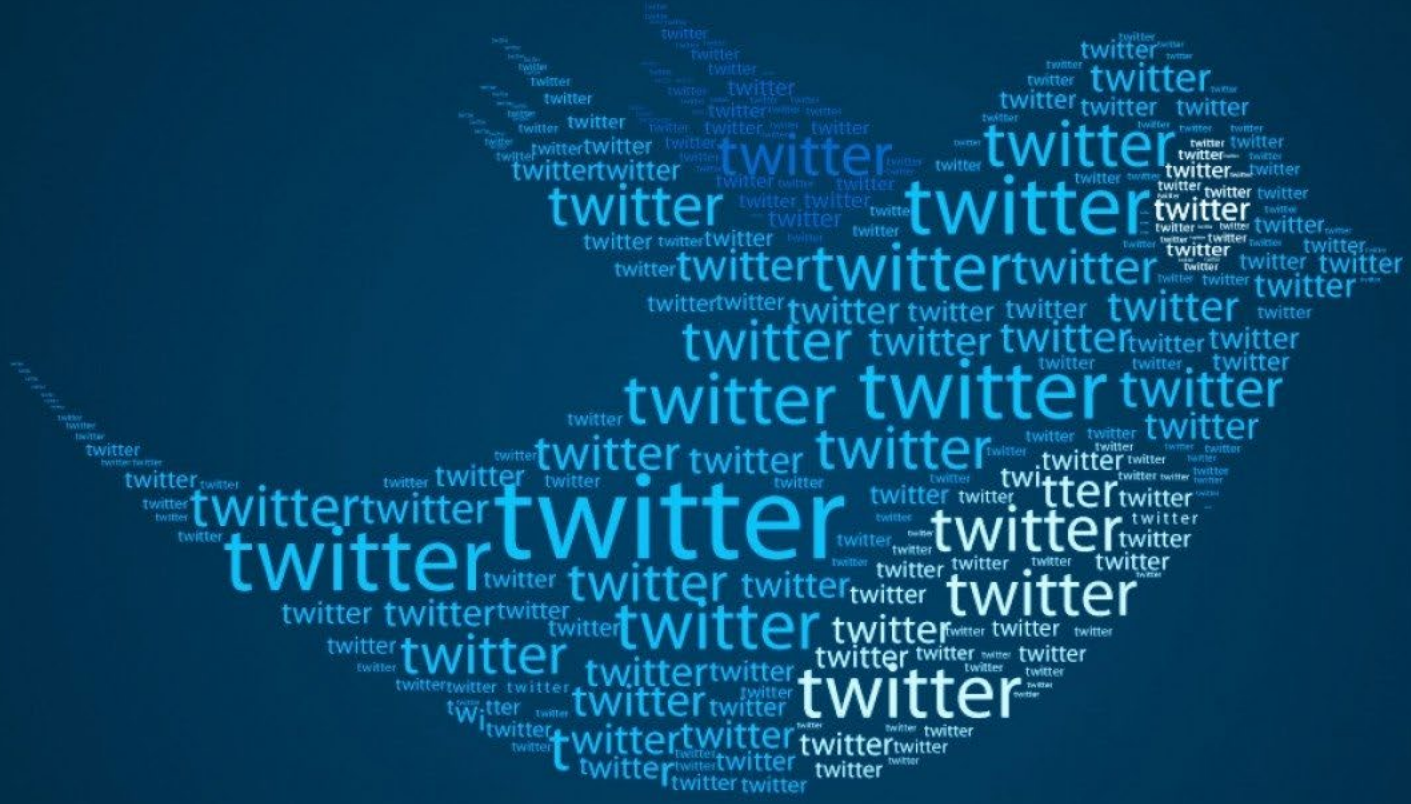
Владислав Бояр
Александра Ивойлова
Мария Мичурина



Определение спама в твитах

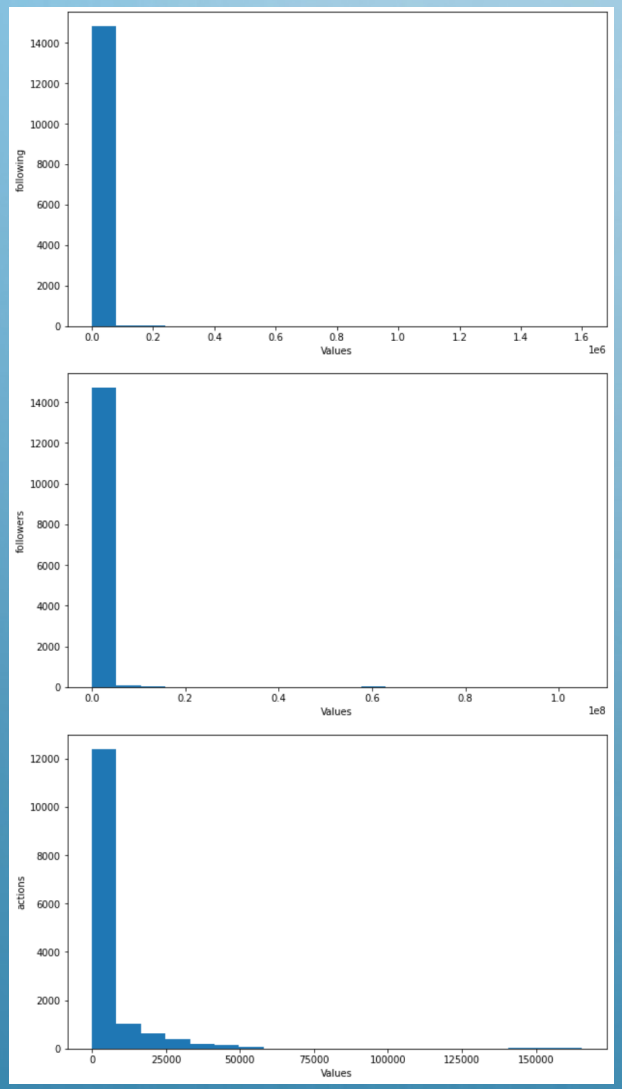
- Задача – классифицировать спам на базе англоязычных твитов
- Обучающая выборка – 14682 уникальных значения
- Имеющиеся данные:
  - собственно твиты
  - количество фолловеров
  - количество читаемых аккаунтов
  - действия (количество произведенных с твитом действий – ретвитов, лайков)
  - локация (есть не у всех)
  - является ретвитом или нет

Feature Engineering

index	Column	Non-Null Count	Dtype
1	Tweet	14899	object
2	following	14741	float64
3	followers	14882	float64
4	actions	11462	float64
5	is_retweet	14898	float64
6	location	12888	object
7	Type	14899	object

- Есть NA в таблице;
- Было обнаружено, что в целевой переменной есть два неправильных значения (South Dakota вместо spam или quality)

Feature Engineering



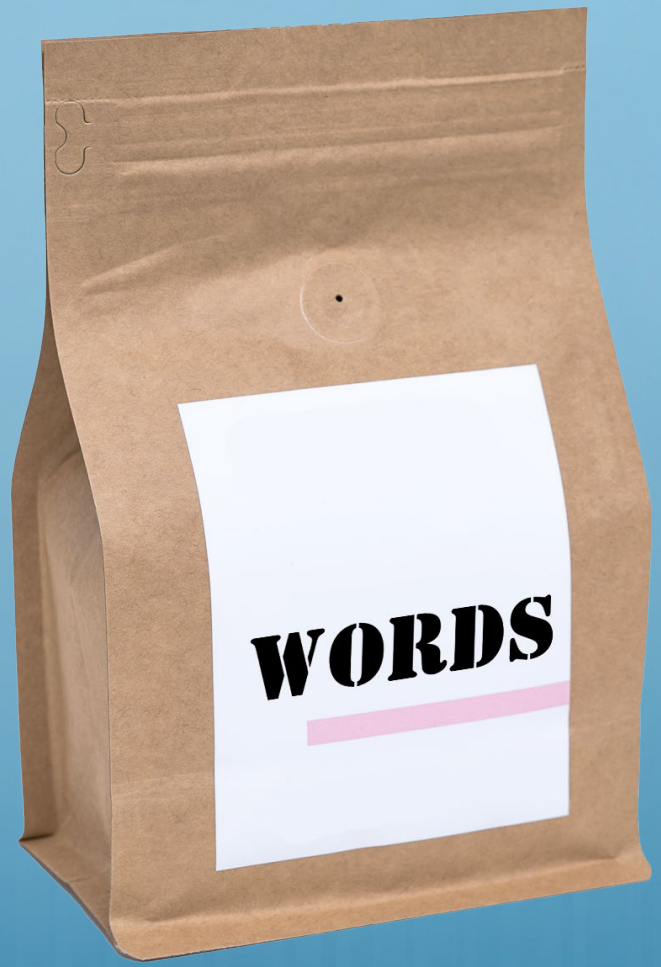
Гистограммы числовых признаков ничего особенно не дали:

- нет нормального распределения;
- большинство значений равно нулю

Feature Engineering. Текстовые признаки

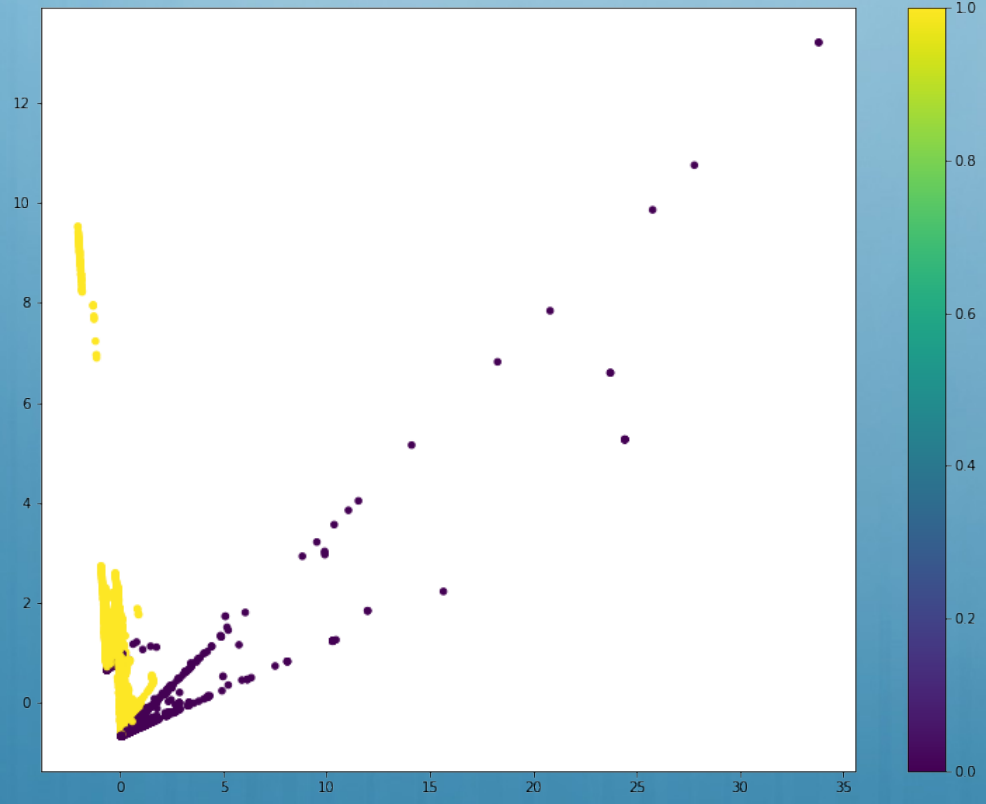
Варианты обработки:

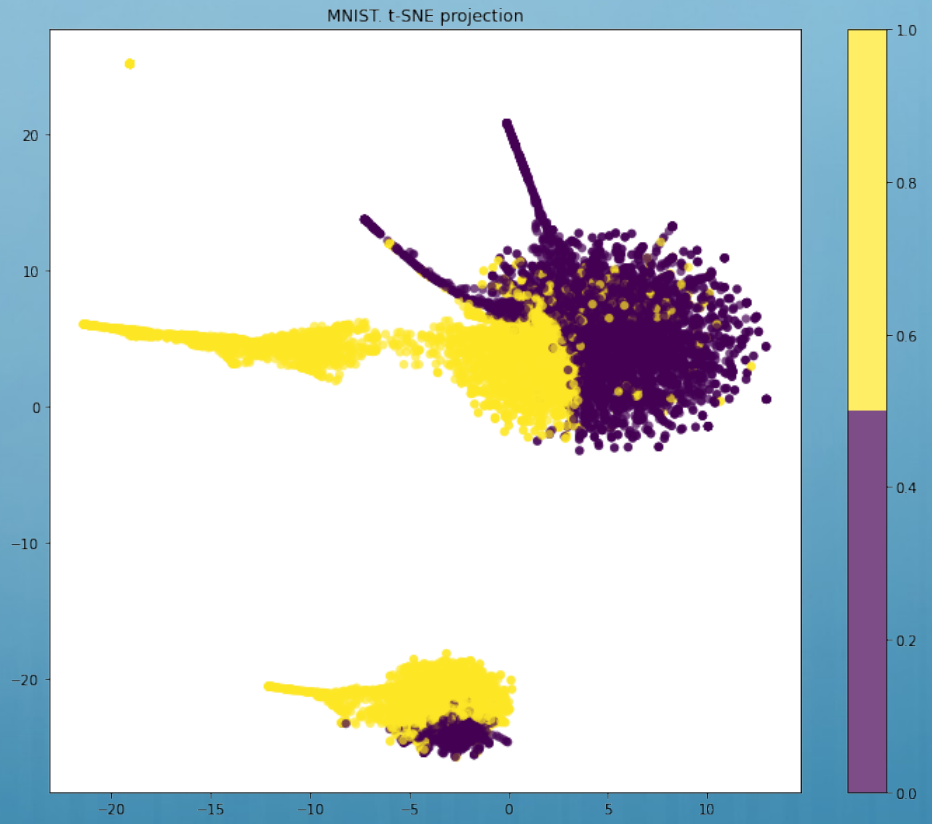
- ✓ Bag of Words
- ✓ TF-IDF
- ✓ unigrams
- ✓ bigrams
- ✓ trigrams



Feature Engineering

Использовали PCA (TruncatedSVD) и t-SNE для визуализации данных; PCA показывает, что данные очень четко кластеризуются. t-SNE ничего внятного не показывает, но красивое





Использованные модели и параметры

- ✓ BoW vs TF-IDF
- ✓ unigrams vs bigrams vs trigrams

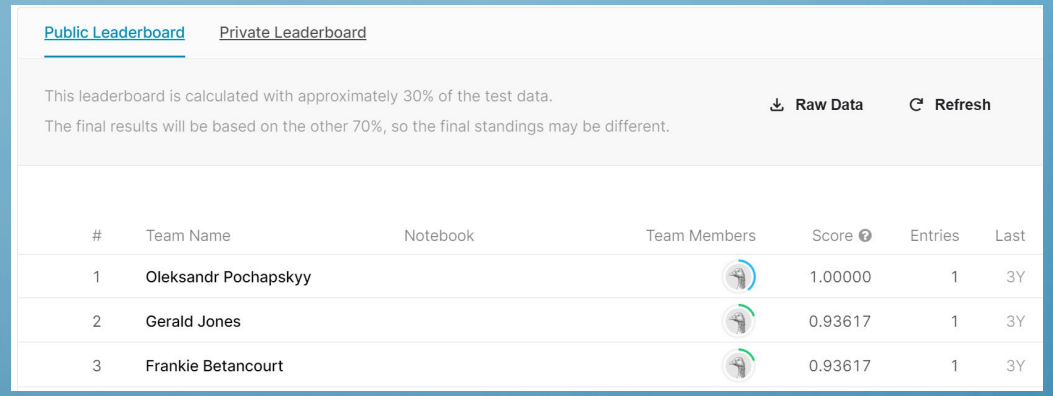
- ✓ Logistic Regression
- ✓ SVC
- ✓ Decision Tree Classifier
- ✓ Bagging Classifier
- ✓ Random Forest Classifier

Результаты

	train (all/only tweets)	dev (all/only tweets)	test (all/only tweets)
BoW, LogReg	0.994 / 0.992	0.956 / 0.895	0.936 / (0.885/0.927)
BoW, SVM	0.984 / 0.976	0.957 / 0.893	0.957 / 0.894
BoW, Tree	0.998	0.994	0.965
BoW, Forest	1.00	0.996	0.982
Tf-idf, LogReg	0.978 / 0.98	0.965 / 0.89	0.936 / 0.876
Tf-idf, SVM	0.98 / 0.997	0.968 / 0.9	0.936 / 0.906
Tf-idf, Tree	0.996	0.982	0.957
Tf-idf, Forest	1.00	0.996	0.97
	2grams only tweets (text)		
	train	dev	test
BoW, LogReg	0.999	0.822	0.851
Tf-idf, LogReg	0.994	0.841	0.78
	2grams		
	train	dev	test
BoW, Forest	1.00	0.997	0.995
Tf-idf, Forest	1.00	0.996	0.995
	3grams		
	train	dev	test
BoW, Forest	1.00	0.997	0.991
Tf-idf, Forest	1.00	0.996	1.00

Победитель:

Random Forest Classifier + TF-IDF + 3-gram



Анализ результатов

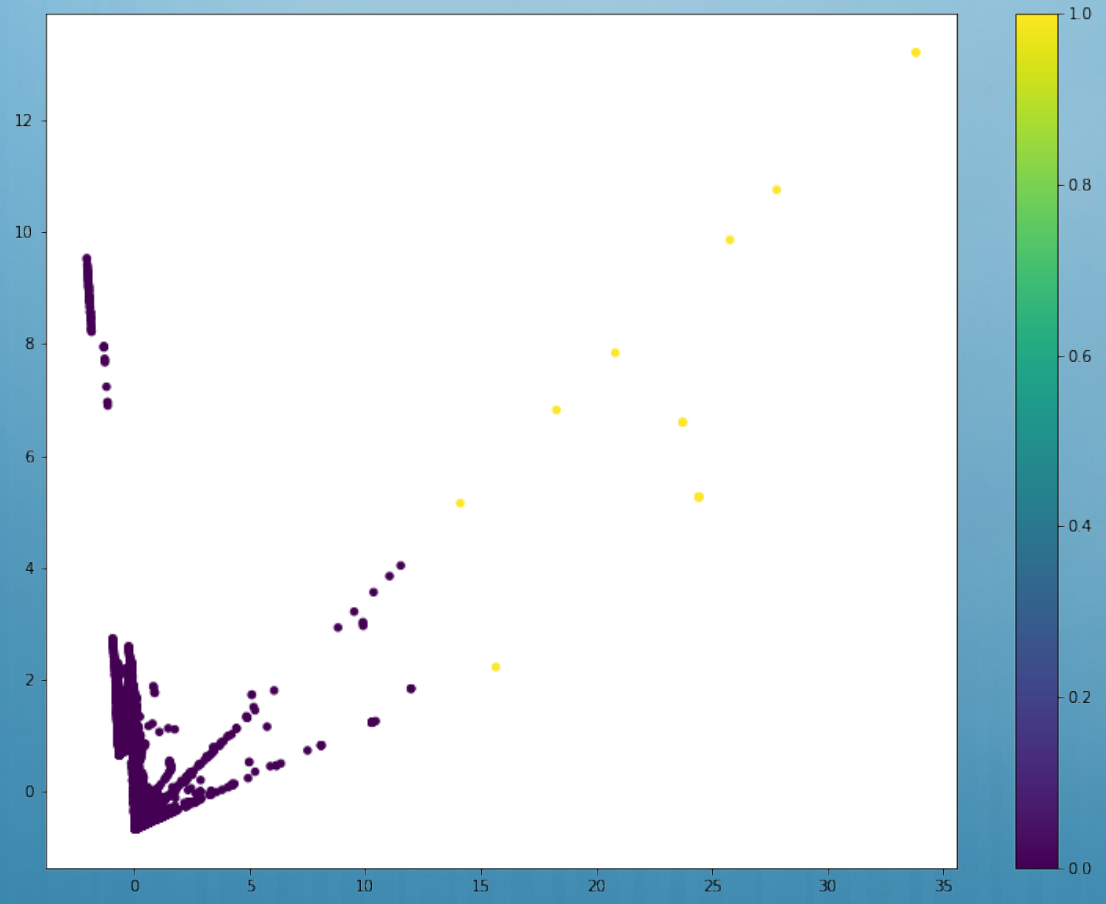
Посмотрели в характеристиках vocabulary_items(), сильнее всего повлиявшие на результат, и обнаружилось, что это ссылки.
Решили проверить:

- собрали все твиты с ссылками в массив;
- посчитали мусорные

Спам со ссылками	Не спам со ссылками
4551	67

Кластеризация (just for lulz)

Agglomerated Clustering



K-Means