Deephound

Первичные исследования классификации постов

11 мая 2017

Манифест разметки

Семантический портрет угрозы

1. Цель сообщения
2. Варианты подачи
3. Первичный словарь

Подготовка к разметке

Этапы подготовки разметки:

1. Беседа со специалистом предметной области
2. Первичная разметка

Формулировка определения и целей угрозы

Угроза 1 – пост, для которого характерен **формат предложения продать** дебетовые карты (разово, оптом, комплектом) или **формат желания купить** дебетовые карты.

Цель Угрозы 1 – осуществить сбыт имеющихся карт, разово продать или купить определенного вида карты, сделать рекламу услуг, наладить постоянный поток сделок по продаже/покупке карт.

Варианты подачи

➢ «куплю/покупаю/покупка...». Четкое предложение о покупке карты определенных банков (желание купить)

➢ «спрос на дебетовые карты...» (желание купить)

➢ «продам/продажа...». Предложение о продаже карт в формате: вид карт, банки, условия, цены, контакты (желание продать)

➢ «нужна/интересует карта...» (желание купить)

➢ «разово/единично продам/куплю/...» (желание разово осуществить сбыт или покупку)

Первичный словарь

Блок глаголы: *«Куплю», «покупаю», «покупка». «Продам», «продаю», «продажа». «Изготовим», «изготавливаем». «Приобрести». «Предлагаем», «предлагаю».*

Блок названия банков: *«Сбер» («сбербанк» и вариации), «Бинбанк» («бин банк»), «промсвязьбанк», «россельхозбанк», «Хоум кредит банк», «райфайзен» банк, «втб», «приватбанк», «открытие», «тинькофф» («тиньков», «тинькоф» и вариации), «альфа» банк.*

Блок видов карт: *«Золото» («голд»), «платина», «классика» («классик»), «виза» («VISA»), «моментальная» («моменталка»), «русский стандарт», «премиум», «кукуруза» и т.д.*

Блок существительные: *«разово», «адекватный», «несколько», «в наличии», «наличие».*

Условия отнесения поста классу Угроза 1

Пост подходит под сформулированное определение

Подача происходит одним из способов описанных подач

Пост имеет словарные единицы из первичного словаря, характерные угрозе

Соотношения классов в выборке

Количество постов	Угроза 1 («+1»)	Others(«-1»)	Доля объектов «+1»
100	37	63	0.37
200	105	95	0.52
300	140	160	0.47
400	164	236	0.41
500	204	296	0.41
600	239	361	0.40
700	268	432	0.38
800	297	503	0.37
900	328	572	0.36
1000	338	662	0.34
1100	364	736	0.33
1200	395	805	0.33
1300	423	877	0.33
1400	439	961	0.31
1500	444	1056	0.30

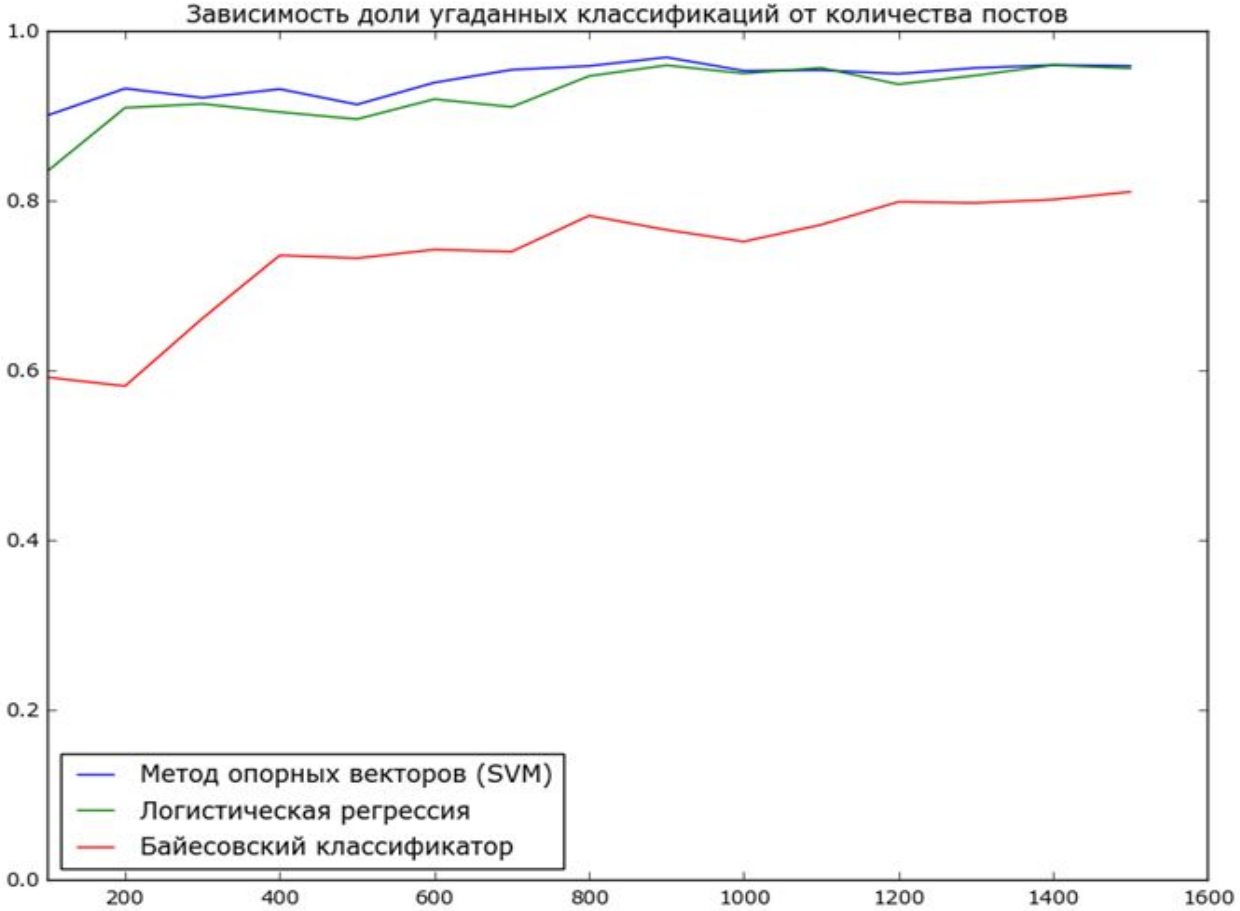
Примитивная предобработка и векторизация

1. Оставляем только русские слова
2. Убираем числа
3. Убираем все символы
4. TF-IDF
5. n-граммы

Самые важные, по мнению классификатора признаки

1. голд 2. куплю 3. продам 4. альфа 5. классик 6. сбер 7. наличии 8. банк 9. карта 10. куплю карты 11. карты 12. продам карты 13. шт 14. втб 15. куплю дебетовые 16. сбера 17. сбербанка 18. куплю карту 19. комплект 20. лс 21. гарант 22. куплю дебетовые карты 23. моментальная 24. разово 25. продам карту 26. доставка 27. полный 28. момент 29. карты приватбанка 30. постоянной

Критерии качества классификатора (Accuracy)



Критерии качества классификатора

TP — истино-положительное решение;

TN — истино-отрицательное решение;

FP — ложно-положительное решение;

FN — ложноо-трицательное решение.

	y(x)=+1	y(x)=-1
a(x)=+1	TP	FP
a(x)=-1	FN	TN

Критерии качества классификатора

Точность:

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP}$$

Полнота:

$$\mathrm{Re}\,call = \frac{TP}{TP + FN}$$

F:

$$F = (1 + \beta^2)\frac{\mathrm{Pr}\,ecision * \mathrm{Re}\,call}{\beta^2(\mathrm{Pr}\,ecision + \mathrm{Re}\,call)}$$

Результаты соревнований методов

SVM

	accuracy	precision	recall	f1	amount
0	0.901170	1.000000	0.722381	0.836580	100
1	0.932698	0.977500	0.890152	0.925563	200
2	0.921803	0.965474	0.858582	0.907982	300
3	0.932025	0.990909	0.814500	0.891804	400
4	0.913838	0.947368	0.814357	0.872326	500
5	0.939681	0.979286	0.861941	0.916520	600
6	0.954762	0.972319	0.899913	0.934239	700
7	0.959211	0.964226	0.921766	0.942039	800
8	0.969439	0.971557	0.948999	0.959430	900
9	0.953612	0.964871	0.873905	0.917023	1000
10	0.954258	0.959726	0.891068	0.923803	1100
11	0.950000	0.958532	0.882447	0.918317	1200
12	0.956963	0.965106	0.901107	0.931661	1300
13	0.960224	0.960748	0.896436	0.927280	1400
14	0.959394	0.966574	0.892981	0.927933	1500

Логистическая регр.

	accuracy	precision	recall	f1	amount
0	0.835673	1.000000	0.549762	0.703846	100
1	0.910000	0.943137	0.884470	0.903636	200
2	0.914605	1.000000	0.816158	0.897023	300
3	0.904924	0.990476	0.735356	0.837973	400
4	0.896611	0.932055	0.778730	0.844788	500
5	0.920149	0.977743	0.811689	0.886453	600
6	0.910971	1.000000	0.755845	0.859958	700
7	0.947368	0.975603	0.879305	0.924798	800
8	0.960055	0.973849	0.921518	0.946543	900
9	0.950459	0.972280	0.856618	0.910557	1000
10	0.957125	0.975758	0.884408	0.927290	1100
11	0.937719	0.956146	0.844978	0.896631	1200
12	0.948040	0.966619	0.871587	0.916165	1300
13	0.960226	0.966290	0.891000	0.926788	1400
14	0.956432	0.978095	0.886384	0.921754	1500

Байес

	accuracy	precision	recall	f1	amount
0	0.592398	0.600000	0.129365	0.208889	100
1	0.582063	0.676154	0.688770	0.536213	200
2	0.661845	0.916667	0.209760	0.338577	300
3	0.736135	0.898644	0.411376	0.548061	400
4	0.732819	0.889123	0.363098	0.511655	500
5	0.742873	0.822276	0.444521	0.564462	600
6	0.740545	0.887818	0.336354	0.485171	700
7	0.782895	0.861810	0.530946	0.644767	800
8	0.766164	0.789057	0.396178	0.521159	900
9	0.752381	0.772923	0.309875	0.438094	1000
10	0.772139	0.760899	0.429027	0.543777	1100
11	0.799123	0.770178	0.583194	0.654887	1200
12	0.797884	0.753632	0.473280	0.578685	1300
13	0.801816	0.738121	0.535142	0.614754	1400
14	0.810936	0.741470	0.499573	0.593051	1500

F метрика


Зависимость f1-критерия от количества постов
1.0
0.8
0.6
0.4
0.2
0.0
200
400
600
800
1000
1200
1400
1600
SVM
Logistic Regression
BernouliNB

Развитие DEEPHOUND. Система классификаторов