



# TO FIND OUT OR TO BUY? PRODUCT REVIEW VS. WEB SHOP CLASSIFIER

**Р**avel Braslavski

**У**ri Kiselev

Dialog 2011

# What does user actually mean?

Яндекс

Нашлось  
2 млн ответов

алые паруса

в найденном  в Екатеринбурге

- a novel by Alexander Grin?
- its screen version of the 1961?
- a residential complex in Moscow?

# Who is behind the query?

Яндекс

Нашлось

2 млн ответов

fixed assets amortization

в найденном  в Екатеринбурге

- experienced accountant
- college student

# The diversity of search results

When it is impossible to disambiguate the query, we can try to organize the result list the way it reflects the different intents of the query.

# Skew of a search engine result list

Яндекс

Нашлось  
52 млн ответов

ноутбуки

в найденном  в Екатеринбурге

 [Ноутбуки на Маркете](#)

[HP](#) [ASUS](#) [Acer](#)

Популярные модели. Описание и отзывы. Сравнение характеристик и цен.  
[market.yandex.ru](#) > Ноутбуки

1  [Купить Ноутбуки Екатеринбург, компьютеры Екатеринбург, мониторы...](#)

[Ноутбуки](#) [Компьютеры](#)

**Ноутбуки** (под заказ) 319. Планшетные устройства 3. Компьютеры 399. ... Время работы в праздничные дни!! 26 Апреля 2011 **Ноутбук** HP! Два ядра!

 [Екатеринбург, ул. Восточная, 11б](#) [все адреса](#) +7 (343) 262-46-65

[MKcomputer.ru](#) [Екатеринбург](#) [копия](#) [ещё](#)

2  [интернет магазин компьютерной техники](#)

Доставка: Екатеринбург

Найден по ссылке: Заказывайте на сайте [pro100good.ru](#): интернет магазин **ноутбуков** Заходите.  
[pro100good.ru](#) [ещё](#)

3  [Noutika.ru. Продажа ноутбуков в Екатеринбурге - доставка ноутбуков...](#)

Доставка: Екатеринбург

Количество **ноутбуков**, участвующих в акции ограничено. **Ноутбуки** бесплатно доставляются по г. Екатеринбург, а также в г. В. Пышма и в г.... подробнее.

[noutika.ru](#) [Екатеринбург](#) [копия](#) [ещё](#)

4  [НЭТА - Товарный каталог](#)

**Ноутбуки** мультимедийные. Коммуникаторы и КПК. Электронные книги. Моноблоки. Планшетные ПК. Комплектующие для **ноутбуков** и КПК.

[neta.ru](#) > Каталог [Екатеринбург](#) [копия](#) [ещё](#)

5  [www.space97.ru | Главная | Продажа компьютеров и ноутбуков Asus, Sony...](#)

[Ноутбуки](#) [Нетбуки](#) [Планшеты](#) [Асус](#) [Ленув](#) [Самсунг](#) [МПК](#)

# Product queries

query: [*samsung g400*]

query: [*home air conditioner*]

query: [*asus motherboard*]

query: [*lg monitor*]

query: [*gps navigation systems*]

# Main intents

- comparison, studying the properties of goods
- purchase
- technical documentation
- software for devices
- etc.

# Web documents classifier

Two types of documents:

1) online surveys and **reviews**;

2) pages of **online shops** you can immediately make an order on.



# Online shop

**EKBSHOP**  
интернет-магазин  
электроники

Контакты  
Тел.: (343) 219-17-88, 8-908-920-8888  
ICQ: 176272  
Email: ekbshop@mail.ru

Ваша корзина  
Товаров: 0 шт.  
Всего к оплате: 0 руб.

О магазине    **Доставка и Оплата**    Гарантия    Статус заказа    Контакты

ПОИСК:  Выберите категорию  от  до

- Телекоммуникационное оборудование
- Сотовые телефоны и коммуникаторы
  - Apple iPhone
  - Коммуникаторы ASUS
  - Коммуникаторы Glofish Eten
  - Коммуникаторы HTC**
  - Сотовые телефоны LG
  - Сотовые телефоны Motorola
  - Сотовые телефоны Nokia
  - Сотовые телефоны Philips
  - Сотовые телефоны Samsung
  - Сотовые телефоны Sony Ericsson
  - Сотовые телефоны Toshiba
  - Bluetooth USB-адаптеры
  - Bluetooth гарнитуры
  - Проводные гарнитуры
  - Зарядные устройства для телефонов
  - Аккумуляторы для телефона
- Аудио и Видео
- Автомобильные устройства
- Видеокамеры
- Мини-АТС и факсы
- Карты Памяти



## Коммуникаторы HTC купить в Екатеринбурге!

Новейшие **коммуникаторы HTC** от именитого производителя, зарекомендовавший себя исключительно с лучшей стороны. Разработчики заменили привычный джойстик на так называемый джог-болл - миниатюрный вариант трекболла. HTC оснащены не только всеми беспроводными модулями (от телефонного GSM/GPRS/EDGE до Bluetooth и Wi-Fi), но и GPS-приёмником.

1 2 >> | [показать все](#)



Коммуникатор HTC A3333 Wildfire grey

- Экран TFT 3,2", 16777216 цв.
- 64-голосная полифония
- MP3-мелодии
- Камера 5 Мрх
- Bluetooth, USB, Wi-Fi
- Размеры 107 x 60 x 13 мм

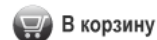
9 500 руб.



Коммуникатор HTC A3333 Wildfire red

- Экран TFT 3,2", 16777216 цв.
- 64-голосная полифония
- MP3-мелодии
- Камера 5 Мрх
- Bluetooth, USB, Wi-Fi
- Размеры 107 x 60 x 13 мм

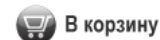
9 500 руб.



Коммуникатор HTC A3333 Wildfire white

- Экран TFT 3,2", 16777216 цв.
- 64-голосная полифония
- MP3-мелодии
- Камера 5 Мрх
- Bluetooth, USB, Wi-Fi
- Размеры 107 x 60 x 13 мм

9 500 руб.



# Review document

- Intel Centrino 802.11 a/b/g/n wireless
- Bluetooth 2.1+EDR
- 90W Smart AC adapter with HP Fast Charge
- 6-cell (62 WHr) Li-Ion battery
- Weight: 6.05 lb
- Dimensions (w x d x h): 14.72 x 9.86 x 1.34 inches
- MSRP: **\$1,499.00** (starting price: \$1,099.00)



## Build and Design

In recent years the HP EliteBooks have distinguished themselves in the business world thanks to an exterior design featuring brushed-metal cladding and an interior chassis made of durable magnesium alloy. In short, the EliteBooks look cool and are built tough. The latest generation of HP EliteBooks takes that heritage a few steps further

with what HP calls its "FORGE" design philosophy. FORGE is actually an acronym for the words, "Form, Optimization, Richness, Green and Enduring." If you want to overlook that marketing fluff for a moment, what HP is trying to say is that these notebooks are stylish, offer excellent performance, deliver a premium feel, provide efficient power management for long battery life, and are very well built to survive the rigors of business use.

Those might sound like bold claims but, based on what we've seen from the previous generation of EliteBooks, those marketing claims have a very real basis in fact.

One of the first things you'll notice when you pick up the new 15-inch EliteBook 8560p is that the notebook feels like it's made of very thick chunks of aluminum. You won't feel the "flex" that you normally see in cheaper plastic laptops. The



# Classifier

To compose a three-class classifier, we built two binary classifiers: “*shop – other*”, “*review – other*”.

## Classifier:

- high performance → light-weight features
- LIBSVM

# Learning sample

Query list	
1	samsung g400
2	monitor LG
3	ручная газонокосилка
	...
110	кофеварка Delonghi

Top10


Поиск: [Почта](#) [Карты](#) [Маркет](#) [Новости](#) [Словари](#) [Блоги](#) [Видео](#) [Картинки](#) [ещё](#)


Яндекс  
Нашлось 789 тыс. ответов

samsung g400  
в найденном в Екатеринбург [расширенный поиск](#)

Результаты [все](#) [в рунете](#) [в мировом интернете](#)

- [Mobile-review.com Обзор GSM/UMTS-телефона Samsung G400](#)  
По характеристикам, конечно же. Возьми в руки Samsung U900 Soul и Samsung G400 - ничего общего между аппаратами нет, эстетика различна.  
[mobile-review.com](#) · [review@samsung-g400.shtml](#) [копия](#) [ещё](#)
- [Купить SAMSUNG G400 SOUL дешевле, сравни цены на САМСУНГ G400...](#)  
О том, как купить Самсунг G400 Soul дешевле. На текущий момент в интернет магазинах нет предложений SAMSUNG G400 SOUL в России и на Украине.  
[mobiq.ru](#) > [Сотовые телефоны и цены](#) > [Samsung](#) > [samsung\\_g400\\_soul.html](#) [Москва](#) [копия](#) [ещё](#)
- [Евросервис - ... коммуникаторов, iPhone - г Екатеринбург - Samsung - G400](#)  
ремонт Samsung G400 Самсунг в Екатеринбурге, белый экран, дисплей Samsung, сенсорный центр Samsung, разблокировка, замена дисплея, код телефона, клавиатуры, шлейфа, ремонт динамика...  
[remtel66.ru](#) > [article/a-1010.html](#) [Екатеринбург](#) [копия](#) [ещё](#)
- [SAMSUNG SGH-G400 SGH-G400XDASER](#)  
SAMSUNG SGH-G400 Сама элегантность G400 - первый в мире телефон с внешним сенсорным дисплеем. Просто коснитесь дисплея, чтобы включить камеру 5 Мпикс, радио, музыкальный или видеоплеер.  
[samsung.com](#) > [ru/consumer/mobile...\\_hnp...G400XDASER...](#) [копия](#) [ещё](#)
- [Мобильный телефон Samsung SGH-G400 Soul. Обзоры описания тесты](#)  
Обзоры, тесты: 27.02.2008 Samsung SGH-G400 Soul: душа цвета металла. 23.07.2008 Обзор GSM/UMTS-телефона Samsung G400 04.07.2008 Samsung G400 Soul review: Fold and touch.  
[helpix.ru](#) > [Samsung](#) > [g400\\_soul](#) [копия](#) [ещё](#)
- [Samsung SGH-G400 - описание, характеристики, тест, отзывы, цены, фото](#)  
Информация на сайте о Samsung SGH-G400: обзор, тест, отзывы, купить, сравни цены, продать Samsung SGH-G400 на форуме, описание с характеристиками и фото.  
[zoom.onews.ru](#) > [goods\\_card/character...samsung..g400](#) [копия](#) [ещё](#)
- [Samsung SGH-G400 сотовый телефон - MobiSet.Ru](#)  
Честно говоря, давно дизайн мобильного телефона не вызывал у меня таких эмоций, какие вызвал Samsung G400. Мохет, а просто осознучило по раскладушкам?  
[mobiset.ru](#) > [Каталог мобильных телефонов](#) > [mobile/?id...](#) [копия](#) [ещё](#)
- [Samsung SGH-G400. Каталог мобильных телефонов - Сотвик](#)  
Одной из главных особенностей Samsung G400 является наличие внешнего полностью сенсорного дисплея. В дополнение к внутреннему 2,22-дюймовому TFT ЖК-дисплею Samsung G400 оборудован...  
[sotovik.ru](#) > [Каталог](#) > [Samsung](#) > [SGH-G400 Soul](#) [копия](#) [ещё](#)

«samsung g400» в картинках  
  
[Все картинки](#)

Видео «samsung g400»  
  
Samsung G400 Soul Review (in Roman) - [www.TelefonuTau.eu](#)  
[Все видеоролики](#)

Assessor

Labeled set

Shop / Review / Misc



# Learning sample – shop classifier

<b>Class</b>	<b># of pages</b>
Shop	301
Review	87
Misc	591
Total	979

# Learning sample – review classifier

<b>Class</b>	<b># of pages</b>
Review	150
Misc	150
Long docs	50
Total	350

# Test sample

Class	# of pages
Shop	431
Review	101
Misc	557
Total	1089

The test sample was obtained the same way as the shop training sample.

# Classification features

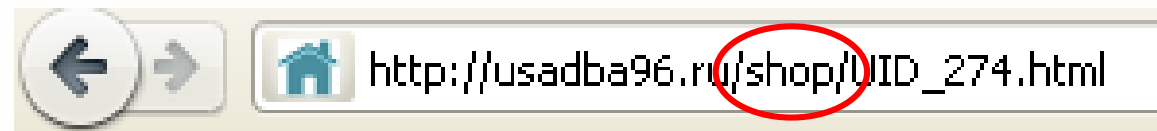
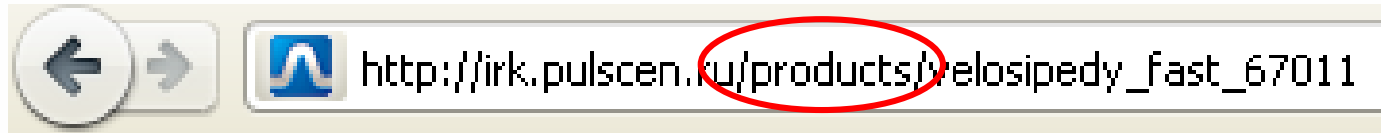
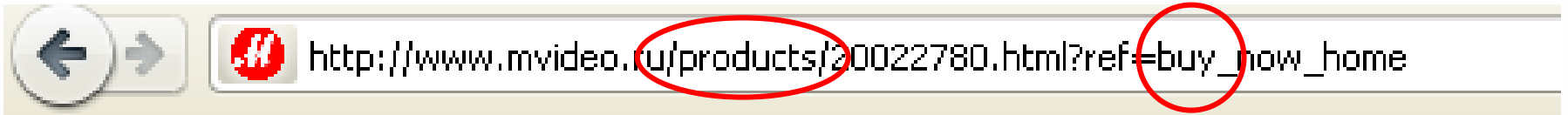
- ***Term features.*** The most contrast terms for shop documents are магазин, рубль, каталог, цена, прайс, and корзина (*shop, ruble, catalog, price, and basket*).
- ***HTML features.*** The possibility to make an order (“buy” button detection)



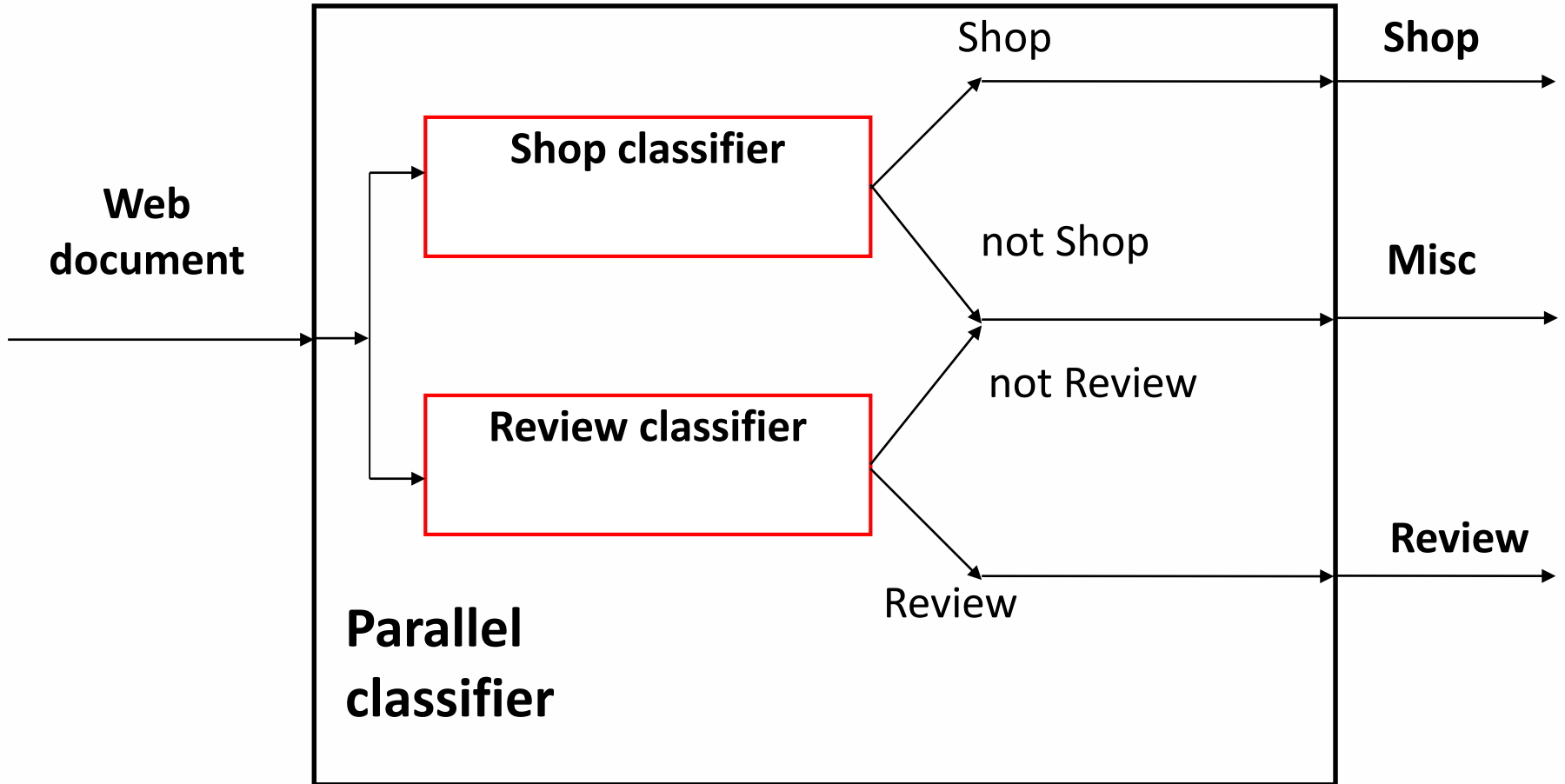
# Classification features - 2

- **Lexical features.** Dictionary of trademarks and brands (**Sony, Samsung, Asus**) presented in Yandex.Market service + *appraising adjectives* (**стильный, ужасный, прекрасный, отвратительный** – *stylish, awful, wonderful, terrible*)
- **Textual features.** The length of document in words and characters, distribution of sentence lengths, and etc.

# URL features



# Parallel classifier



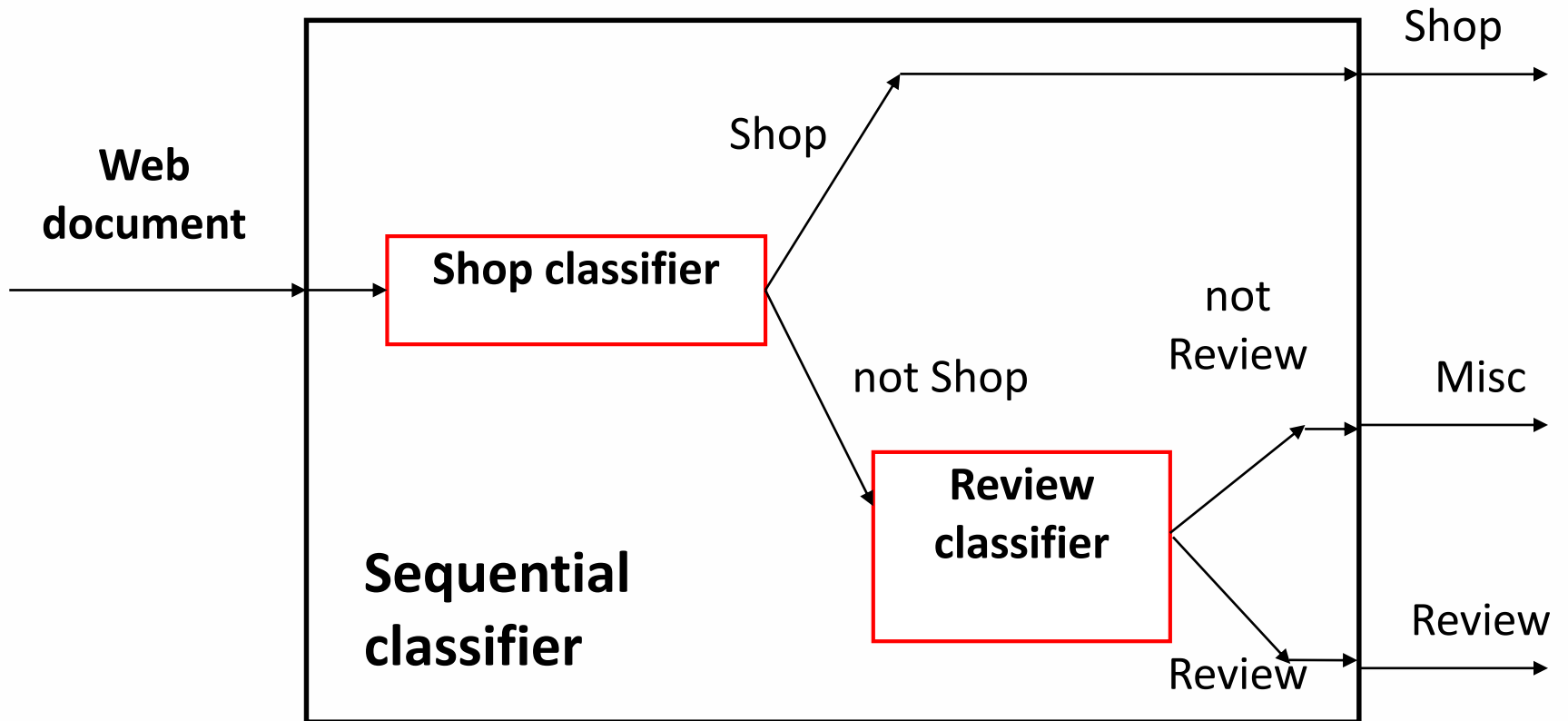
# Shop classification results

Set of features	Precision	Recall
Terms only	0.918	0.809
HTML features only	0.894	0.491
Term + HTML features	0.934	0.800
Term + lexical features	0.910	0.807
Term + URL features	0.876	<b>0.856</b>
All features	<b>0.937</b>	0.837

# Review classification results

Set of features	Precision	Recall
Terms only	0.644	0.861
Term + URL features	0.643	0.841
Term + lexical features	0.625	0.861
Term + textual features	<b>0.681</b>	<b>0.891</b>

# Sequential classifier



# Confusion matrix of the three-class classifier

	Shop	Review	Misc	Recall
Shop	361	3	67	<b>0.84</b>
Review	1	90	10	<b>0.89</b>
Misc	23	23	511	
Precision	<b>0.94</b>	<b>0.79</b>		

# Conclusion and future work

- Light-weight features
- Small learning samples
- Acceptable quality of classification
- High performance

**We want to improve the review classifier:**

- Offline classification
- Add more sophisticated features





**P**avel Braslavski (pb@yandex-team.ru)

**Y**uri Kiselev (yurikiselev@yandex-team.ru)