

Opis scenariusza

- WVCorp — firma zatrudniająca Ciebie (analityka danych).
 - Firma WVCorp dysponuje forami tematycznymi i dyskusyjnymi dla każdego produktu, gdzie klienci mogą dyskutować o problemach i cechach.
 - “Szum medialny” — wtedy, gdy jakiś temat na forum użytkowników ma bardzo duży poziom aktywności, co wskazuje na zainteresowanie użytkowników tym zagadnieniem.
- eRead — czytnik e-booków wyprodukowany przez firmę WVCorp.
- TimeWrangler — aplikacja do zarządzania czasem wyprodukowana przez firmę WVCorp.
- BookBits — czytnik e-booków wyprodukowany przez konkurencję.
- GCal — stworzona przez inną firmę rozproszona usługa kalendarzowa, którą można zintegrować z aplikacją TimeWrangler.

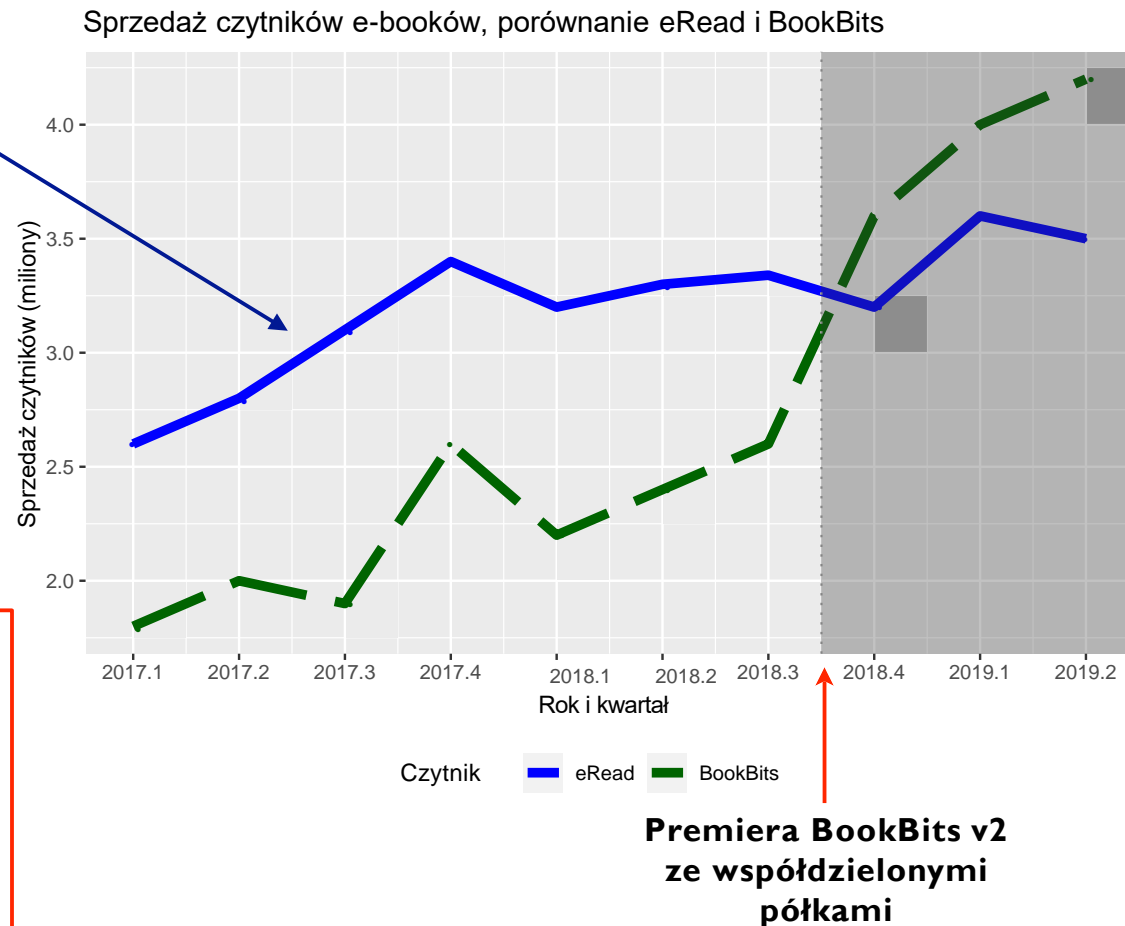
Przewidywanie szumu medialnego

Hipotetyczna prezentacja
dla sponsora projektu
przygotowana przez zespół
analityków danych WVCorp

Stracona okazja

eRead — wcześniej
najlepiej sprzedający
się czytnik e-booków.

Szacunkowo
utracono 25
mln dolarów
przychodów
ze sprzedaży
produktu.



Czy mogliśmy to przewidzieć?

- Dyskusje na forum produktu eRead:
 - Współdzielenie listy ulubionych książek ze znajomymi.
 - Udostępnienie książki grupie znajomych (kto pierwszy, ten lepszy).
- Zawsze, gdy pojawiały się pytania na ten temat, dyskusja ożywała.
 - Sugestie, obejścia, prowizorki, duża zgodność wśród klientów.
 - Wspólna półka (jak w BookBits) spełniałaby te wciąż powtarzane prośby.
- Wokół tego tematu narastał **szum medialny!** Ale zignorowaliśmy go. Albo przeoczyliśmy.
 - Ciągłe obserwowanie aktywności na forum jest bardzo pracochłonne.

Cel: wczesne wykrywanie

- Przewidywanie, które tematy na forach produktu będą cechować się ciągłym szumem medialnym.
- Cechy wyczekiwane przez klientów.
- Istniejące cechy, które sprawiają problemy.
- Trwały szum: rzeczywiste, trwające potrzeby klientów.
- Problemy ani chwilowe, ani wynikające z trendów.

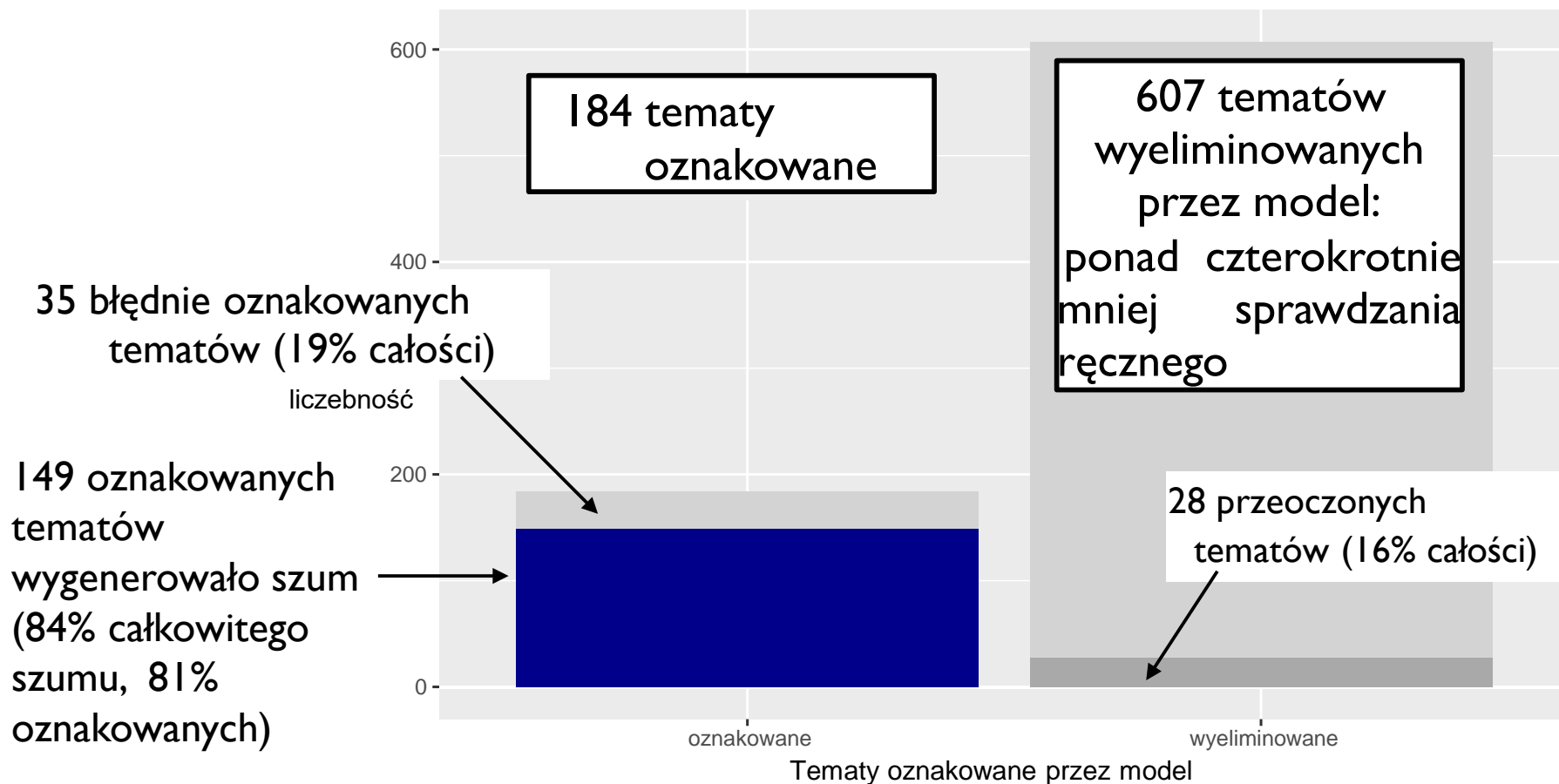
Badanie pilotażowe

- Zebrano na forum dane z trzech tygodni.
- W pierwszym tygodniu wytrenowano model, aby sprawdzić, które tematy będą szumieć medialnie w dwóch kolejnych tygodniach.
- **Szum = utrzymywany wzrost 500+ aktywnych dyskusji na dany temat w ciągu dnia w porównaniu z pierwszym dniem w pierwszym tygodniu.**
- Przewidywane rezultaty porównano z rzeczywiście popularnymi wątkami.
- Opinie od pięcioosobowego zespołu menedżerów produktu — jak przydatne były wyniki?

Obiecujące wyniki

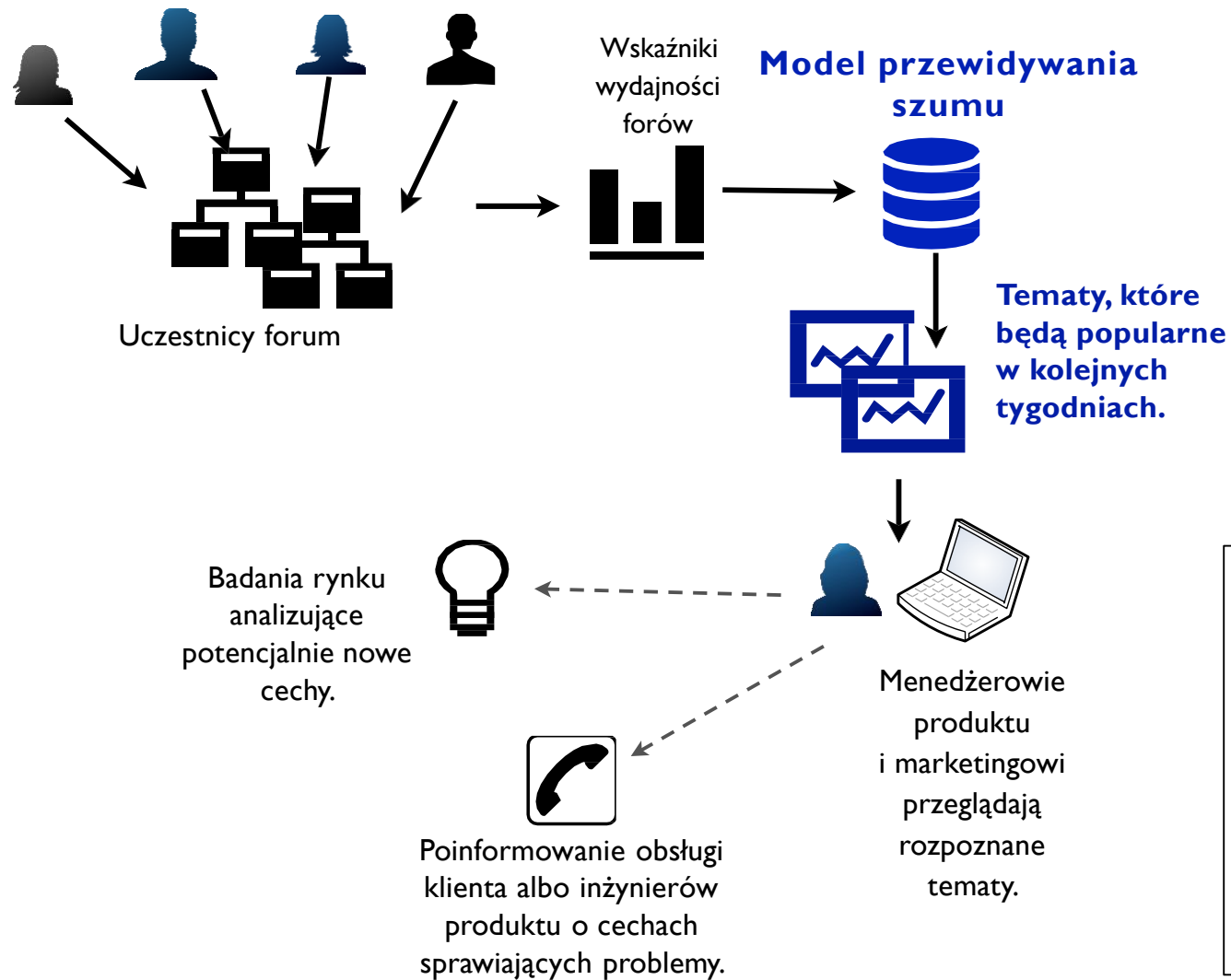
Wyniki wykrywania szumu medialnego

Całkowita liczba tematów: 791; 177 wygenerowało szum medialny



Klienci: 75% rozpoznanych tematów dało “wartościową opinię”.

Mechanizm działania



Korzystamy z już zaimplementowanych wskaźników:

- Liczba autorów w danym temacie.
- Liczba dyskusji w danym temacie.
- Liczba wyświetleń tematu.
- Itd.

Model Buzz

- Model lasu losowego.
 - Głosuje wielu „ekspertów”.
 - Działa wydajnie na dużych zestawach danych.
 - Obsługuje dużą liczbę zmiennych wejściowych.
 - Kilka ogólnych założeń dotyczących interakcji zmiennych lub ich istotności.
 - Bardzo dokładny model.

Przykład: wczesne wykrywanie problemu

- Temat: TimeWrangler → Integracja z GCal
 - Liczba dyskusji wzrosła od czasu opublikowania GCal v7.
 - Zdarzenia GCal niespójnie wykrywane; niewłaściwy mechanizm ich oznaczania.
 - Zadania TimeWrangler przechodzą do złej aplikacji Gcalendar.
- **„Gorąco” na forach, jeszcze zanim zrobiło się „gorąco” w dziennikach zdarzeń obsługi klienta.**
 - Aktywność na forum pobudziła model dwa dni po aktualizacji GCal.
 - Obsługa klienta spóźniona o tydzień.

Co dalej?

- Dalsze zmniejszenie obciążenia menedżerów produktu, przekazywanie im lepszych informacji od klientów.
 - Nowe wskaźniki poprawiające przewidywania.
 - Rejestrowanie wzrostu/spadku aktywności dyskusji (a także jego szybkości).
 - Dlaczego rejestrują się nowi użytkownicy na forum? Jakiego zadają pytania?
 - Cel: Wykrycie 98% nieuchronnego szumu przy 10% zdarzeń fałszywie pozytywnych.
- Skuteczne przekazywanie informacji o szumie do właściwych menedżerów produktu, marketingowych i grup obsługi klienta.

Dziękuję