Automatyczne wykrywanie fałszywych treści i pomiar abstrakcji w języku

Aleksander Wawer

21 stycznia 2013

axw@ipipan.waw.pl

Współautorzy: Maciej Rubikowski, Dominika Rogozińska, Katarzyna Bitka

Plan referatu

1. Opinion Spam (źródła problemu)
2. Metody rozpoznawania zmyślonych i nieprawdziwych treści
3. Pomiar abstrakcji językowej za pomocą Linguistic Category Model
4. Korpus fałszywych recenzji w języku polskim i eksperymenty

Spam i nieprawdziwe recenzje

Internauta o pseudonimie **observer2010** na stronie cokupic.pl (serwis należący do Grupy Allegro) umieścił jak dotąd 24 opinie, z czego 23 to oceny pralek Electroluksa. Wszystkie entuzjastyczne. Przykładowe nagłówki: *Odlotowa pralka*, *Bardzo ciekawa pralka*, *Piękny sprzęt do prania*, *Czołówka ekstraklasy* itd. Końcowa ocena - zawsze najwyższa, pięć gwiazdek. Co ciekawe, **observer2010** jednego dnia potrafi przetestować aż cztery pralki! Wszystkie 24 testuje między lipcem 2009 a styczniem 2010 r.

Źródło:
`http://m.wyborcza.biz/biznes/1,106501,11277293,Jak_firmy_kupuja_przychylne_recenzje_w_internecie_.html`

Metody wykrywania spamu wśród opinii

Komentarz **wyciorex**

[..] marketingowcy piszą ogólnikowo, i często śmieszą mnie ich komentarze, bo ich komentarz można z powodzeniem wstawić do każdej branży, to samo z negatywami, jak czytam, że totalny złom bo się zepsuła już kilka razy - bez jakiś szczegółów, to wiadomo, że konkurencja. Jak się komuś coś zepsuje to napisze, że w e52 przestał działać głośnik, lub w laptopie urywa się klapa matrycy.

Metody wykrywania spamu wśród opinii

Istnieje wiele metod nie opartych w ogóle o analizę językową lub mocno korzystających z innych źródeł informacji niż język:

- Wykrywanie grup fałszywych recenzentów (Mukherjee et al., 2012): czas zamieszczenia opinii, liczba i charakter tych opinii (np. wszystkie entuzjastyczne).
- Analiza liczby produktów, wspólnie ocenianych przez grupę recenzentów (Mukherjee et al., 2011).

Cechy językowe kłamstwa

Za Newman et al. (2003), nieprawdziwa komunikacja (*deceptive communication*) ma następującą charakterystykę:

- użycie dłuższych wypowiedzi, dłuższych zdań i *“bigger words”*. Elokwencja i nienaturalne, dalekie od potocznych sformułowania ("zabetonować czas prania");
- mniej zaimków osobowych, zwłaszcza pierwszej i trzeciej osoby liczby pojedynczej;
- więcej słów o negatywnym wydźwięku (ściślej, negatywnych emocjach);
- więcej czasowników oznaczających ruch (motion verbs).

Typy czasowników w LCM

Jak mierzyć abstrakcyjny język?

- Czasowniki Stanów (SV) odnoszą się do stanów mentalnych i emocjonalnych, kognitywnych (myśleć, rozumieć, itd.) lub afektywnych (nienawidzieć, podziwiać, itd.). SV nie odnoszą się do pojedynczych obserwowalnych zdarzeń, ale reprezentują trwałe, przedłużające się stany, które nie mają jasnego początku ani końca.
- Deskryptywne lub Interpretatywne Czasowniki Działań (DAV/IAV) odnoszą się do określonych działań (np. uderzyć, pomóc, plotkować, itd.). Działania te mają jasno określony początek i koniec.

Typy czasowników w LCM

- Kryterium wyróżniającym DAV (Deskryptywne Czasowniki Działań) jest to, że posiadają one przynajmniej jedną powiązaną z nimi cechę fizyczną.
  - Przykładowo, *całować* implikuje usta jako fizyczną cechę, *chodzić* - nogi, itd.
- Pierwszym kryterium wyróżniającym IAV jest odniesienie do wielu różnych działań, które mają to samo znaczenie, ale nie mają wspólnej cechy fizycznej. Są zazwyczaj neutralne (wydźwięk), ale mogą zyskać nacechowanie w określonym kontekście.
  - Przykładowo, *pomagać* możemy przeprowadzając staruszkę przez ulicę lub pożyczając koledze pieniądze.
- Drugim kryterium IAV jest wyraźny komponent ewaluatywny (nie neutralny wydźwięk). Przykładowo, *pomagać* lub *zachęcać* są pozytywne, *oszukiwać* lub *znęcać* są negatywne.

LCM w języku angielskim

The General Inquirer, słownik Harvard IV (Stone et al., 1965)

Wielowymiarowy (183 kategorie) słownik z 11,768 słów i sensów słów, powiązany z formalizmem ujednoznaczniania sensów (Kelly & Stone 1975). Wśród kategorii są między innymi trzy dotyczące LCM: IAV, DAV, SV.

- 1947 czasowników IAV,
- 540 czasowniki DAV,
- 102 czasowniki SV.

LCM - pomiar abstrakcji

Kodowanie poziomu abstrakcji jako wartość przedziału 1-4 pkt.

- Beavis *uderzył* Butt-Heada. (DAV,1 pkt)
- Beavis *zranił* Butt-Heada. (IAV, 2 pkt)
- Beavis *nie lubi* Butt-Heada. (SV, 3 pkt)
- Beavis jest *agresywny*. (ADJ, 4 pkt)

Źródło: `http://www.cratylus.org` (Semin & Fiedler 1988)

LCM narzędziem pomiaru abstrakcji

- Oczekiwania: Linguistic expectancy bias (Wigboldus et al., 2000)
  - Zachowania oczekiwane: opis abstrakcyjny
  - Zachowania nieoczekiwane: opis konkretny
- Stereotypy: Linguistic intergroup bias (Maas et al., 1989)
  - Zachowania stereotypowe: opis abstrakcyjny
  - Zachowania wbrew stereotypom: opis konkretny
- Wywieranie wpływu na innych (cele komunikacji).

LCM w języku polskim

Tłumaczenie metodą automatyczną z angielskiego

Bing Translation API, tłumaczony napis:
"I can <czasownik>."

- Tokeny kończące się na "ć" (większość) lub "c" (wlec, piec, itd.).
- Wynik: 1183 polskich form czasownikowych w bezokoliczniku z przyporządkowaną informacją o LCM angielskich odpowiedników.

LCM w języku polskim

Ręczna weryfikacja LCM i wydźwięku

- Dwie osoby anotujące (R1 i R2),
- zgodność wobec oryginału (EN).

	Wydźwięk		LCM	
	Zgodność (%)	Kappa	Zgodność (%)	Kappa
R1 vs EN	95.6	0.93	94.7	0.88
R2 vs EN	99.2	0.98	82.2	0.63
R1 vs R2	94.9	0.92	80.1	0.57

Kategoryzacja czasowników w LCM jest trudniejszym problemem, niż kategoryzacja tych czasownikow ze względu na wydźwięk.

Operacjonalizacja w języku polskim - wyniki

Polski LCM. Pierwszy fragment.

Powstał przez automatyczne tłumaczenie czasowników w The General Inquirer i ręczną weryfikację pod kątem LCM (i wydźwięku).

	GI	EN2PL	R1	R2
IAV	1947	828	801	854
DAV	540	226	257	205
SV	102	37	38	42

Jak przypisać LCM pozostałym czasownikom w języku polskim?

Automatyczne przypisywanie kategorii LCM

Czy możliwe jest **automatyczne** określenie, do jakiej kategorii LCM należy dany czasownik?

- Jeżeli tak, to kluczowa jest analiza rozkładów typów rzeczowników, z którymi wiąże się dany czasownik.
- Hipoteza: czasowniki DAV powinny wiązać się z dopełnieniami o rozkładach bliższych fizykalnym i najmniej abstrakcyjnym fragmentom w grafach hiperonimii. IAV i SV powinny być odpowiednio bardziej abstrakcyjne.
- Nie interesują nas rzeczowniki-podmioty zdania, tylko dopełnienia, przedmioty czynności wyrażanej czasownikiem.

Zapytanie w Poliqarp

```
[base=<...>][pos!=subst&pos!=interp]?[pos=subst&case=acc|gen]
```

Automatyczne przypisywanie kategorii LCM

- Dla każdego rzeczownika, zbiór krotek w postaci: (najwyższy hiperonim, odległość od wierzchołka), (..,..), (..,..)
- Na tak stworzonych danych uruchamiamy typową procedurę selekcji cech i uczenia maszynowego.
- Wstępne wyniki są obiecujące, jednak nie można uznać, że klasyfikacja tego typu jest wystarczająco dobra dla stworzenia polskojęzycznego zasobu LCM. Słowosieć 1.8. Zastosowanie Random Forest w 10*CV daje średnią precyzję równą 0.685.

Korpus Szczerości

Celem było stworzenie zasobu oddzielającego ..

- opinie ewidentnie zmyślone (nie wywołane doświadczeniami z ocenianym produktem lub usługą, nie powstałe spontanicznie tj. wywołane) od
- niemal na pewno prawdziwych (powstałych spontanicznie i na skutek doświadczeń z ocenianym produktem lub usługą).

Zasób udostępniony jest na stronie
`http://zil.ipipan.waw.pl/KorpusSzczerosci/`

Korpus Szczerości

- Skąd można wiedzieć, że opinie lub recenzje są fałszywe?
- Paradoksalnie, jest to najprostsze. Wystarczy je kupić (Ott et al., 2011).

Coffeenka ★ 8 (+16)

do negocjacji	do negocjacji	2012-10-09 14:06
cena	termin wykonania	ostatnie zmiany

Witam, bardzo chętnie podejmę się tego zlecenia, mam odpowiednie doświadczenie i umiejętności, aby profesjonalnie je zrealizować. Perfumami, zapachami i aromaterapią interesuję się już od dawna i mam dużą wiedzę tym temacie - m.in. pisałam teksty o tej tematyce dla portalu zapachowy.pl oraz tworzyłam teksty na stronę internetowej perfumerii multiperfumeria.pl Przez kilka miesięcy pracowałam też w internetowym sklepie z kosmetykami, gdzie tworzyłam ich profesjonalne opisy .Pisanie to moja największa życiowa pasja, posiadam "lekkie pióro" i potrafię stworzyć tekst na każdy temat w ten sposób, by był on zrozumiały w odbiorze dla potencjalnego odbiorcy. Z wykształcenia jestem magistrem antropologii kulturowej i absolwentką Międzynarodowego Studium Dziennikarskiego. Od kilku lat zajmuję się pisaniem różnorodnych tekstów. Pracuję sama, dokładnie i terminowo. Bardzo dobrze znam język polski, stosuję aktualne reguły jego gramatyki, ortografii i stylistyki zawarte w najnowszych słownikach. Naturalnie gwarantuję unikalność tekstów oraz zgodność z wymaganiami. Wstępnie proponuję 1 zł za każdy opis żądanej długości, więc 100 zł za całość. Co do terminów dostosuję się do wymogów. Moje artykuły są do wglądu na stronie: http://tekstyaleksandry.blogspot.com/ W razie potrzeby mogę podesłać także inne moje teksty. Proszę o kontakt i szczegóły na: [redacted]@tlen.pl. Pozdrawiam serdecznie i zapraszam do współpracy.

luki-freelancer2010 ★ 9 ★ 2 (+27)

300 PLN	3	2012-10-09 11:36
cena	termin wykonania	ostatnie zmiany

Portfolio: www.twierdzagier.pl. Kontakt: [redacted]@wp.pl, Tel. [redacted], Skype paqo1991. Proszę o komentarz po zakończeniu aukcji lub pisemną referencję. Pozdrawiam z Chełma, [redacted]

Korpus Szczerości

Skąd brać prawdziwe (?) recenzje?

- Opłacenie ich powstania nie ma sensu: prawdziwe recenzje tworzone są spontanicznie, w przypływie chęci podzielenia się opinią o produkcie lub chęci krytyki. Dodatkowo, nie znamy rzeczywistych użytkowników ocenianych produktów lub usług i nie ma możliwości dotrzeć do nich z prośbą o recenzję.
- Wykorzystaliśmy teksty zamieszczane na jednym z forów dyskusyjnych (niech pozostanie Nienazwane), które zawierają także powiązaną z nim bazę recenzji. Do Korpusu Szczerości trafiły tylko recenzje pisane przez osoby, które prowadzą też aktywność na forum dyskusyjnym i nie pisały recenzji na masową skalę.

Korpus Szczerości

Do korpusu trafiło 500 recenzji dwóch typów perfum, najpopularniejszych według Innego Dużego Serwisu: CK Euphoria i D&G Light Blue. Trafiło mniej-więcej po równo recenzji pozytywnych i negatywnych.

- 1 .. 100 – CK Euphoria, autentyczne
- 101 .. 150 – CK Euphoria, fałszywe (coffeenka)
- 151 .. 200 – CK Euphoria, fałszywe (coffeenka – krótsze)
- 201 .. 250 – CK Euphoria, fałszywe (luki freelancer)
- 251 .. 350 – D&G Light Blue, autentyczne
- 351 .. 400 – D&G Light Blue, fałszywe (coffeenka)
- 401 .. 450 – D&G Light Blue, fałszywe (coffeenka – krótsze)
- 451 .. 500 – D&G Light Blue, fałszywe (luki freelancer)

Korpus Szczerości - CK Euphoria

'Euphoria' to seksowny zapach działający na wyobraźnię. Został stworzony dla kobiet, które dążą do absolutnej wolności. Są pewne siebie i charakteryzuje je wrodzona elegancja. To zapach, którym Calvin Klein zaskoczył wszystkich. Świeży, lekko owocowy sprawia, że poczujesz euforyczny przypływ energii, czerpanej z egzotycznych kwiatów i owoców. Flakon, którego inspiracją były pąki orchidei zaprojektował Fabien Baron.

Nuty zapachowe:

- **nuta głowy**: owoc granatu, daktyle, zielona nuta
- **nuta serca**: kwiat lotosu, kwiat champaca, czarna orchidea
- **nuta bazy**: płynna ambra, czarny fiołek, akord kremowy, drzewo mahoniowe.

Korpus Szczerości - D&G Light Blue

Light Blue to kontynuacja klasycznego stylu zapachów Dolce&Gabbana. Kompozycja jest ukłonem w stronę piękna kobiety. Pokazuje jak cieszyć się każdym aspektem życia, afirmuje zarówno małe przyjemności, jak i porywające, gorące uczucia. Kompozycja kwiatowo - owocowa wyrażająca radość życia. Błękitny, aksamitny futerał skrywa kryształowy flakon. Zwieńcza go transparentny korek przepasany srebrną wstążką. Flakon nawiązuje stylistyką do klasycznej kompozycji Dolce&Gabbana pour Femme, jest jej subtelną i świeżą wersją.

Nuty zapachowe:

- **nuta głowy**: jabłko Granny Smith (zielone jabłuszko), cedr sycylijski, dzwonki
- **nuta serca**: jaśmin, biała róża, bambus
- **nuta bazy**: cedr, ambra, piżmo.

Korpus Szczerości - Uczenie Maszynowe

Za (Newman et al., 2003) warto wykorzystać rozkłady części mowy (zwł. zaimków i przymiotników), tagów morfoskładniowych (zwł. osoba), informacje o wydźwięku. Tabela przedstawia średnią precyzję w 5-krotnej walidacji krzyżowej, klasyfikator SVC.

Cechy	Ilość	Selekcja	C	Precyzja
tagi, POS, wydźwięk, LCM, długość	95	RFE->25	1.0	0.835
tagi i POS	88	-	0.01	0.809
wydźwięk, LCM, długość	8	-	0.01	0.711
przymiotniki i LCM	4	-	1.0	0.639

Korpus Szczerości - Wnioski

- Możemy automatycznie odróżniać prawdziwe i fałszywe recenzje z zadowalającą precyzją. Jest to zadanie wykonalne z wykorzystaniem cech pozaleksykalnych, czyli bardziej uniwersalnych i niezależnych od dziedziny. Potwierdza to wyniki prac Newman et al. (2003) oraz Ott et al. (2011).
- Uzyskane wyniki można prawdopodobnie poprawić, stosując metody do innych typów produktów.
  - Recenzje perfum są bardzo specyficzne (najtrudniejsze?), częste jest użycie metafor i języka figuratywnego.
  - Opinie dotyczą głównie cech pozafunkcjonalnych, a omawiane przedmioty nie mają prawie żadnych parametrów mierzalnych.
- W badaniach Ott et al. (2011), średnia precyzja równa 0.898 osiągnięta została z wykorzystaniem zbioru cech opartych o LIWC i bigramy, czyli leksykalnych.