Raport inițial de cercetare

Am primit un corpus de test format din 7 fișiere, conținând texte clasificate manual în 3 categorii:

- **pozitive** (*„favorabil”* sau *„foarte favorabil”*),
- **negative** (*„nefavorabile”* sau *„foarte nefavorabile”*)
- **neutre** (*„nici favorabile, nici nefavorabile”*)

Pe aceste texte am folosit un algoritm de clasificare automată în categorii, ale cărui rezultate le-am verificat apoi cu cele obținute manual. Deși rezultatele obținute au fost aparent destul de bune, din păcate acestea nu sunt foarte relevante.

În primul rând, fișierele de test sunt prea mici pentru a putea fi folosite cu succes în procesul de învățare al algoritmului. Fiecare fișier conținea în jur de câteva mii de intrări (*tabelul 1*), insuficient pentru a stabili niște reguli folositoare. În plus, multe dintre aceste mesaje se repetau, cum este cazul celor provenite de pe Twitter.

În multe situații, intrările din fișiere conțineau doar un mic fragment din începutul textului analizat, cum ar fi în cazul postărilor de pe bloguri. Acestea nu ofereau de multe ori niciun semn al opiniilor exprimate în restul textului, deci nu erau folositoare analizei noastre.

Una dintre cele mai importante probleme a fost faptul ca majoritatea textelor aparțineau categoriei neutre. După cum se poate observa din *tabelul 1*, unde prezentăm rezultatele clasificării textelor în două categorii (neutre și cu opinie), procentele de texte clasificate corect sunt foarte mari, dar destul de irelevante. Ținând cont că marea majoritate a textelor erau neutre, chiar dacă algoritmul nostru le-ar fi clasificat pe toate drept neutre, am fi obținut un procentaj bun.

Mult mai aproape de adevăr este procentajul obținut pentru fișierul cu texte legate de **ursus**, unde raportul celor două categorii era aproape egal. În acest caz au fost clasificate corect 76% din texte, procentajul scăzut datorându-se numărului total mic de texte din care algoritmul putea învăța anumite reguli.

Am încercat și să clasificăm doar textele cu opinie în cele două categorii, pozitive și negative (*tabelul 2*). Raportul dintre clase a fost mai bun, chiar dacă și în acest caz textele pozitive au fost de două ori mai multe decât cele negative, însă numărul total de texte cu opinie a fost foarte mic.

Procentele clasificare corectă au fost destul de mari, în general peste 82%, dar din nou nu foarte relevante. Motivul este că numărul mic de texte face algoritmul să învețe niște reguli foarte specializate, care vor returna rezultate foarte bune când sunt verificate din nou pe acestea. Chiar și în cazul care am combinat toate fișierele au rezultat doar 2101 de texte cu opinie, destul de puțin pentru un astfel de algoritm.

În concluzie, deși algoritmul pe care l-am construit a obținut rezultate bune pe fișierele de test, pentru a putea face o antrenare și o testare cu adevărat relevantă a algoritmului avem nevoie de un număr mult mai mare de texte clasificate manual. Acestea ar trebui să nu se repete și să fie texte întregi, pentru că fragmentele de început s-au dovedit de multe ori irelevante. În plus, un raport echilibrat între textele din cele 3 categorii este esențial pentru obținerea de rezultate corecte.

Keyword	Total texte	Texte neutre	Texte cu opinie	Clasificate corect	Clasificate gresit
bcr	6055	5853	202	97.01%	2.99%
blogjuan	2240	1961	279	91.79%	8.21%
cremosso	343	260	83	84.84%	15.16%
danone	1168	876	292	86.22%	13.78%
fratelli	539	520	19	97.40%	2.60%
upc	1025	570	455	76.20%	23.80%
ursus	3787	3016	771	81.75%	18.25%

Tabelul 1 – clasificare în texte neutre și cu opinie

Keyword	Total texte	Texte cu opinie	Texte pozitive	Texte negative	Clasificate corect	Clasificate gresit
bcr	6055	202	28	174	92.07	7.92
blogjuan	2240	279	220	59	87.81	12.18
cremosso	343	83	63	20	89.15	10.84
danone	1168	292	120	172	82.19	17.8
fratelli	539	19	13	6	57.89	42.1
upc	1025	455	326	129	84.17	15.82
ursus	3787	771	584	187	83.65	16.34
Total	15157	2101	1354	747	82.29	17.7

Tabelul 2 – clasificare în texte pozitive și negative