Rezumat articol 1(Simona):

Articolul se focuseaza pe distinctia dintre limbajul "hateful" si limbajul "offensive". S-au facut teste pe 25k tweets, alese random din postarile a 33458 utilizatori Twitter. Aceste postari au fost prelucrate manual de angajatii CrowdFlower, avand la dispozitie definitia uni limbaj ofensiv si un paragraf explicativ, ca mai apoi ei sa le poata clasifica in una din cele 3 categorii: hate speech, offensive but not hate speech, or neither offensive nor hate speech.

S-au testat mai multe modele, printre care logistic regression, naive Bayes, decision trees, random forests, and linear SVMs. Modelul final la care au ramas este regresia logistica cu regularizare L2.

S-a ajuns la concluzia ca, cel mai mare obstacol este identificarea unui limbaj hateful care nu contine injuraturi sau comentarii rasiste/homofobice si ca este foarte important contextul mesajului.

A fost folosita structura sintactica I <intensity > <user intent > <hate target> pentru identificarea mesajelor hateful.

Rezultatele lor demonstreaza modul in care "hate speech" este folosit, el poate fi folosit direct catre un grup de persoane sau un individ, poate fi folosit exprimand o opinie generala sau poate fi folosit in conversatii dintre persoane.

Rezumat articol 2(Michael):

Articolul arata ca acest topic se imparte in 2 parti: hate speech detection si offensive language detection si incearca sa clasifice din tweet uri, care sunt normale, hateful si offensive.

Pentru acest clasificator se folosesc algoritmi de machine learning: SVM - Support Vector Machines si Deep Neural Networks, Bidirectional Encoder Representations from Transformers.

In jur de 1-2% din numarul total de tweet uri sunt offensive. Procesarea tweet urilor este importanta deoarece apar multe emojiuri si semne de punctuatiie, diacritice, textul nu este neaparat corect gramatical.

Bineinteles ca exista si cateva erori, dar cele mai multe au fost la tweet uri cu referinta la serialul Game of Thrones, citind acele tweet care nu erau offensive, clasificatorul le gasea ca fiind offensive si era gresit, sau gresea pentru ca un cuvant dintr-un tweet era offensive dar tweet ul in sine nu era, ex: cuvantul "steal".

Concluzia articolului este ca chiar si cu algoritmi avansati si abordari de machine learning, mereu vor fi si greseli, si ca pe viitor daca ar exista mai multe date de antrenare probabil ar fi mai bun algoritmul.

Rezumat articol 3(Daniel):

Pentru procesarea limbajului natural in detectia limbajului ofensator mesajele sunt mai intai pre-procesate. Deseori mesajele din social media contin emoji, url-uri, tag-uri. Acestea trebuiesc inlaturate, ca si cuvintele din lista stop-words a limbii romane.

Dupa pre-procesare, setul de date trebuie impartit in 3 parti, unul pentru antrenare, altul pentru testare si un test pentru validare. Daca seturile de date de antrenare si testare nu au label-uri, se poate folosi three-level hierarchical annotation model pentru adnotare.

Urmeaza pasul de tokenizare folosind FastText, Universal Sentence Encoder sau DMD.

Contextul este important pentru categorizarea corecta a mesajelor.

Rezumat Articol 4(Georgiana):

-detectarea hate speech-ului si a limbajului ofensator cu ajutorul algoritmilor de Machine Learning de clasificare.

-datele sunt postari de pe site-uri de socializare(Facebook,Twitter)

-clasificatorii utilizati: liniar, SVM, MLP(perceptron multistrat)

-modelul este alcatuit din 3 sub-task-uri:

A (identifica daca postarea contine hate speech,text profant etc sau nu), B(identifica tipul continutului negativ: hate speech, limbaj ofensator sau limbaj profan),

C(identifica daca postarea are ca tinta un grup/individ sau nu)

-rezultatele sunt bune luand in calcul simplitatea modelului

-SVM este un clasificator mai bun pentru limba engleza, MLP pentru limba germana, in timp ce pentru hindi exista o combinatie echilibrata intre MLP si SVM.

-putem observa ca rezultatele clasificatorilor sunt influentate de numarul de postari preluate pentru fiecare limba si de context.