Въведение

В последните няколко години в Интернет се наблюдава съществен ръст на публикуваните и разпространявани фалшиви новини, които представят изцяло неверни или манипулирани факти (Fake News). Друг (добре забравен) феномен са публикациите с подвеждащи заглавия, които имат за цел да подтикнат Интернет потребителите да отворят линка, като по този начин генерират рекламен трафик (Click Baits). И двете явления имат силно негативен социален ефект, тъй като манипулират общественото мнение и "атакуват" възприятията на потребителите чрез количество, за сметка на качество и достоверност. Класическите журналистически принципи за проверка на фактите, "засичане" на информацията от повече от един независим източник и пр. все още се прилагат от реалните медии, но тъй като спазването на тези принципи струва много време и ресурси, те не могат да компенсират и да насмогнат на несъразмерно по-големия обем фалшиви и подвеждащи публикации.

С масовото навлизане на социалните мрежи (основно Facebook и Twitter), разпространението на подобни публикации до милионни аудитории става все по-достъпно и евтино. Самото съдържание обикновено се "имплантира" в анонимни онлайн сайтове и блогове, като едно и също съдържание се публикува едновременно в десетки различни "източници", които се различават по домейн, структура на страниците, дизайн и пр., но по същество се контролират автоматизирано или полуавтоматизирано от едно и също място. Социалните мрежи се използват предимно като платформи за разпространение на подобни публикации, като за целта се използват фалшиви непроследими профили, които "ударно" споделят подобни публикации в множество страници и групи във Facebook, и през множество фалшиви акаунти в Twitter, Instagram и т.н. Разчита се на т.нар. "Viral Effect". Все по-голяма част от потребителите консумират съдържание от мобилни устройства, реагират първосигнално на провокативни публикации и ги споделят директно, без да се задълбочават в съдържанието, а често и без да са прочели текста отвъд заглавието. Крайният ефект е, че дадена провокативна (респективно фалшива) новина се появява от множество различни "източници" и създава впечатление у хората, че "…то това е по всички новини…".

И двата феномена (Fakes News / Click-Baits) не са нови явления – те съществуват, от както съществуват медии. Новото са каналите за разпространение и лесният и евтин достъп до аудитория, без географски и други физически ограничения. Преди масовото навлизане на Интернет и социалните мрежи в ежедневния живот на хората мащабът и възможностите за генериране и разпространение на фалшиви новини винаги е бил естествено ограничен до тиража на "жълтите" издания или аудиторията на отделни електронни медии. Създаването на стотици анонимни медийни "брандове", които да публикуват едни и същи фалшиви новини, би било непосилна задача във физическия свят, както от финансова, така и от логистична гледна точка, докато в Интернет това вече е факт.

Скандалите в последната една година, свързани с манипулирането на общественото мнение чрез "фалшиви новини" преди президентските избори в САЩ (2016) и Франция (2017) разкриват само част от проблема. Същите техники и подходи се прилагат успешно и за кампании за "черен ПР" срещу компании и неправителствени организации, а в контекста на държавната пропаганда, и срещу цели общности ("мюсюлмани", "бежанци", Европейски съюз, НАТО и пр.).

Предизвикателствата

Балансът между "свободата на словото" и "слободията на словото"

Постигането на баланс е изключително трудна задача – от една страна свободата на словото е човешко право, от друга страна публикуването на неверни (измислени) факти е клевета, и се преследва от закона. Проблемът е, че фалшивите новини се публикуват от анонимни източници и

се разпространяват чрез фалшиви (анонимни) профили в социалните мрежи, а и дори да не са напълно анонимни, времето и ресурсите за завеждане и водене на съдебно дело за всеки отделен случай на предполагаема клевета са несъразмерно по-големи от времето и ресурсите необходими за генериране на фалшиво съдържание и измислени факти. Цензурирането и блокирането на анонимни източници, което се практикува в държави като Китай, Русия и Турция, противоречи на основни демократични принципи, а и на практика не е особено ефективно.

Социалните мрежи и техния корпоративен интерес

В последните месеци компании като Facebook, Google и Twitter направиха официални изявления, че предприемат конкретни мерки срещу разпространението на фалшиви новини. Facebook обяви, че ще работи в партньорство с няколко световни медии за журналистическа проверка на факти, като същевременно назначава допълнително 3000 човека, които да разглеждат и проверяват сигнали за неподходящо съдържание и фалшиви профили в социалната мрежа. Важно е да се отбележи, че и двата подхода включват ангажирането на допълнителен човешки ресурс - въпреки че Facebook е един от лидерите в "Machine Learning" и анализа на данни, все още не всеки проблем може да бъде решен с алгоритми. Основната причина е, че проверката на факти изисква проактивни човешки действия, които не са по силата на алгоритми. Ето два примера:

- Проверка, дали даден "източник" е реално съществуващ, дали наистина е направил дадено изявление, както и дали изявлението е извадено от контекст, или е публикувано обективно и без манипулация. Такъв тип журналистическа проверка изисква идентифициране на източника, оценка на неговата достоверност и разговор/кореспонденция за проверка на обстоятелствата и фактите.
- Проверка, дали даден сигнал за фалшив профил е основателен или злонамерен – докладването за "фалшив профил" във Facebook е лесно, но често пъти съвсем легитимни и влиятелни потребители биват блокирани автоматично, защото името им съдържа псевдоним/прякор, докато реално фалшиви профили, не могат да бъдат блокирани, защото отговарят на определени формални правила (акаунт проверен с мобилен номер, профилна снимка, реални имена и пр.)

Тепърва ще бъдат оценявани резултатите мерките, които предприемат компании като Facebook, но трябва да се има предвид, че социалните мрежи са изправени пред същите предизвикателства и рискове, пред които са изправени и органите на законодателната и изпълнителната власт на държавно ниво – балансът между "свобода на словото" и "цензура". Въвеждането на регулации в Интернет, независимо дали на държавно или корпоративно ниво, води съответно до политически рискове за държавите и юридически/финансови рискове за компаниите. В контекста на социалните мрежи, корпоративните интереси не могат да бъдат пренебрегнати – Facebook печели от количеството генерирано и рекламирано съдържание (директни приходи от реклама), а така също и от броя и ръста на регистрираните и активни потребителски профили в социалната мрежа (индиректно, през цената на акциите). Няма официални статистики или изследвания, каква част от профилите във Facebook са фалшиви, но техният дял вероятно е съществен и масовото им блокиране може да повлияе негативно на цената на акциите и обезценка на компанията – последното нещо, което акционерите биха поощрили.

Има ли просто решение на проблема, има ли "сребърен куршум"?

По всичко личи, че просто решение няма. Има различни решения, за различни аспекти на проблема с фалшивите новини. Мерки, които включват допълнителни човешки усилия за проверка на факти и източници, няма как да осигурят достатъчна мащабируемост, предвид мащабите на Интернет, и

няма как да изключат субективния фактор, поради което е малко вероятно да постигнат съществен и масов резултат.
В крайна сметка основната "тежест" остава при крайните потребители на съдържание, които преценяват, на какво да вярват, дали и до колко да проверяват източниците и информацията, която възприемат и дали да споделят тази информация със своите приятели и последователи.

Основна цел на Хакатона
Целта на настоящия Хакатон е да "атакува" проблема от различна посока, а именно да се експериментира с "Machine Learning" алгоритми, които отговарят на следните условия:

1. Да могат да бъдат "обучени" с помощта на ръчно подготвен и оценен обучителен корпус от публикации.
2. Да могат да идентифицират или оценят вероятно фалшиви / подвеждащи новини само на база заглавия и съдържание, и евентуално на база сходство (fuzzy/similarity match) на заглавия и съдържание в рамките на дискретни периоди от време.
3. Да могат да се приложат в "Интернет мащаб" върху десетки милиони публикации на ден – т.е. максимално икономични от към CPU, RAM и IO ресурси.
4. Да могат да се използват за създаването на безплатна обществено-полезна услуга или онлайн инструмент, който да е достъпен за крайните потребители (консуматори на съдържание), независимо от техническата реализация (плъгин за браузър, мобилно приложение и пр.), с идеята да се улесни преценката на потребителите, дали дадено съдържание е подозрително или не (много).
5. Да бъде с лиценз, съвместим с изискванията за отворен код на Open Source Foundation

Обучителен корпус
Генериран е обучителен корпус, на база извадка от публикации в български онлайн медии и блогове:

- Публикациите са оценени ръчно от студенти по журналистика (по-долу е представена информация за оценяването).
- Предвид големия ежедневен обем публикации и ограничения човешки ресурс, периодът за който е генерирана извадката за обучителния корпус е ограничен до един (1) ден (1000+ публикации).
- Броят публикации включени в извадката от всеки източник е пропорционален на общия брой публикации на източника за предходния, съпоставени с общия брой публикации за предходния месец от всички наблюдавани източници, но не по-малко от една (1) публикация (т.е. всеки източник, който е включен в извадката, има поне по една (1) публикация, независимо, дали делът му в месечния обем публикации е под 0.1%).
- При генерирането на набора от данни са изключени спортните източници.
- Корпусът е разделен на три (3) приблизително равни части, като всяка част е дадена за оценяване на трима (3) различни оценители, така че всяка публикация да получи по три (3) субективни оценки.

Информация за направеното оценяване на обучителния корпус:

- В таблицата с оценките има предвидени две колони за оценяване (оранжева антетка):
  - fake_new_score
  - click_bait_score
- Оценките варират от 1 до 3:
  - 1 = Legitimate / Normal
  - 2 = Neutral

    - 3 = Fake / Click Bait
- Оценяването е извършвано само на база на заглавие и съдържание. Поради технически причини, за някои публикации в съдържанието присъства и името или домейна на източника, но това може да бъде подвеждащо, тъй като дори откровените "Fake News" източници публикуват и реални новини, за да балансират микса и да придадат известна достоверност на съдържанието, което публикуват.

Дефиниция за "Fake News":

- Невероятни „факти“, теории на конспирацията, емоционални провокации (страх, омраза, отвращение)
- Съдържа спекулативни / манипулативни твърдения, няма посочени източници
- Съдържа неологизми / квалификации / прилагателни със силно емоционално внушение

Дефиниция за "Click Bait":

- Заглавието е непълно / подвеждащо / разминава се със съдържанието (например заглавието представя „факт“, а в съдържанието няма факти)
- Представената „информация“ е „пикантна“, скандална, адресираща низки страсти ☺

Тестов корпус

Генериран е тестов корпус, на база извадка от публикации в български онлайн медии и блогове:

- Корпусът е генериран на база публикациите за период от +/- един (1) ден от деня, за който е генериран обучителния корпус.
- Корпусът съдържа XXX XXX публикации.