Извличане на топ тенденции от дискусиите на български език в Twitter

Борис Банков

Extracting top trends from Twitter discussions in Bulgarian

Boris Bankov

Abstract

Социалните мрежи предлагат множество възможности за изследване на потребителски мнения и настроения, чрез обработката на съобщенията, които се публикуват онлайн. Тези мнения често съдържат разнообразна, интересна и полезна информация за широк спектър от теми. Поради големия обем от данни и тяхната неструктурирана същност се налага прилагането методи за извличане на знания. В настоящото изследване се представя подход за извличане на топ тенденции от дискусии на български език в платформата Twitter, който е базиран на групиране и анотиране на текстови данни чрез клъстеризация.

Keywords: twitter text mining, text clustering, social media, data mining, bulgarian text mining, bulgarian text clustering

Въведение

Интегрална част от всекидневието ни се концентрира в общуването с останалия свят посредством Интернет. Новите технологични възможности, големите скорости на обмен на данни, разпространението и достъпността на платформите за социално взаимодействие, предразполагат към развитие на алгоритмите, методите и подходите за обработка на големи масиви от информация. В теоретичната постановка на Big Data три са основните характеристики, който представляват предизвикателство пред софтуерните инженери в областта: размер, скорост на поява и обмен и разнообразие на данните. Социалните мрежи несъмнено са един от актуалните първоизточници на развлекателна и неформална информация за обществото, в което живеем. Чрез платформите се разпространяват мнения, новини, рекламират се продукти, обявяват се промоции, коментират се резултати от избори и спортни събития. Twitter привлича значителна част научни и научно-приложни разработки в сферата на анализа на чувства и извличане на мнения, автоматизирано разпознаване на естествени езици, създаване на модели за предсказване на реални събития и др. Интересът към платформата се увеличава постоянно, поради достъпа до неструктурирани и структурирани данни – потребителските съобщения (наречени tweets или туитове) и метаданните, придружаващи ги. Съобщенията са неформални, кратки, често съдържат служебни символи без възможност за лингвистична интерпретация. Като естествен трябва да се разглежда процесът по публикуването на десетки коментари ежедневно от отделните потребителите в социалните мрежи. Интересен проблем представлява автоматично определяне на контекста и темата или предмета на дискусия в конкретно съобщение. Характерът на информационните масиви, извлечени от социалните мрежи, предполага специфични подходи за анализ на текст. С настоящото изследване се разглежда възможността за извличане на туитове, в момента на тяхното публикуване, прави се оценка на алгоритмите на Twitter за разпознаване на български език в съобщенията и се извеждат най-дискутираните теми в платформата за периода от първи до десети Октомври (01.10.2017 – 10.10.2017г.)

1. Теоретични основи на изследването

Интересът към процеса и приложението на извличането на знания нараства експоненциално през последните години най-вече поради факта, че се увеличава количеството дигитална текстова информация, която се публикува в социалните мрежи. Twitter е една от най-често използваните платформи за провеждането на експерименти в областта на извличането на знания от данни и по конкретно извличане на знания от текст.

Платформата предлага възможност за извличане на съобщения по два начина: достъп до архив от публикувани назад във времето туитове (historical data) или чрез абониране към постоянна емисии от данни (online stream) на случващи се в реално време туитове. Съобщенията могат да се филтрират по различни параметри, като един от тях е език. За да се извлекат топ тенденции в дискусиите на български език е нужно първо правилно да се идентифицира използвания естествен език.

В началото на десетилетието научните публикации и разработки се фокусират върху естеството на туитовете, по-конкретно тяхната кратка и неформална структура (Ellen, 2011; Xue et al, 2011; Yin 2013). Съобщенията в Twitter могат да съдържат до 140 символа, като през месец Октомври на 2017-та година започва да се разглежда идеята за удвояване на позволения лимит. Правилното разпознаване на естествен език в подобни кратки текстови конструкции (микротекст) представлява известно предизвикателство. Туит съобщенията, съдържат малко на брой думи, което затруднява стандартните езикови класификатори. В статия, публикувана от екипа от разработчици на Twitter[1] е представен алгоритъмът, който се прилага за автоматично разпознаване на естествения език в даден туит. При изследването на текст, съдържащ формален изказ, понякога е достатъчно да се създаде списък с най-често използваните думи от всеки език. В социалните мрежи изказът се различава от този на новинарските сайтове или този, който може да срещнем в книга, учебник или документ. Допълнителен проблем представлява смесването на езици в рамките на едно съобщение, съкращения и използване на абревиатури и хаштагове, които по-често са на английски език. Ако бъдат разгледани статистически данни за ползваемостта на платформата на различни езици може да се срещне категорията „und" или "undefined" (неопределен). Това обикновено са:

- съобщения, в които се среща комбинация от повече от един естествен език;
- съобщения с произволна последователност от символи;
- съобщения, написани на език, който не може да бъде разпознат (съществуващ, но изключително слабо представен в световен мащаб);
- съобщения, съдържащи единствено изображения и линкове.

В Таблица 1 са представени примерни съобщения, чиято езикова принадлежност трудно може да се определи.

Таблица 1. Примерни съобщения с неопределена принадлежност към естествен език.

№	Текст	Език	Обяснение
1	<3 пица @ Posto di pasta	Неопределен	Съдържа име на място на италиански, думата пица може да е на български или сръбски
2	#hi #pooper #bird #swallow #лястовица	Неопределен	Съдържа думи на английски, но последната е на български
3	Pozor pozor ManU	Неопределен	Може да е чешки, сръбски, харватски, български и т.н.
4	хахаах ʃつ◕_◕ʅつ	Неопределен	Не съдържа смислен текст, може да е руски или български

Алгоритъм първоначално е апробиран с помощта на човешко анотиране върху примерно множество от данни. Лингвисти разглеждат туитовете като на първа стъпка отбелязват езиците, които разпознават с абсолютна сигурност, докато останалите ги

[1] Evaluating language identification performance, < https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance.html>, 19.10.2017

отбелязват като „най-близък или вероятен“ език. В комбинация с това се разглеждат профилите на потребителите, тяхна геолокация и ако е нужно се използват други ресурси като речници например. След това се сравняват резултатите от човешкото анотиране и програмния алгоритъм за разпознаване на език. От резултатите се разбира, че при произволна извадка повечето езици нямат добро представяне в множество. Това означава, че ако се срещнат малко на брой туитове на български не е възможно да се определи точността на алгоритъма. Едно от възможните решения е да се определят по 1000 туита на всеки език от лингвистите за създаването на т.нар. балансирано множество, но това представлява затруднение, туй като съобщенията трябва да бъдат търсени ръчно. Сравнително по-лесно е да се разгледат потребители, които публикуват туитове на определен език. След това се взима множество от съобщения за определен период от време, публикувани от съответните потребители и с помощта на човешко анотиране се премахват всички туитове, които не са на желания изследван език. Измерването на прецизността на алгоритъма върху балансираното множество може да даде заблуждаващи резултати. В изследване на Marco Lui и Timothy Baldwin[2] се стига до заключение, че алгоритъма на Twitter за разпознаване на естествени езици не произвежда съществени разлики спрямо свободно достъпни средства като библиотеката langid.py за Python, Compact Language Detector 2 на Google, LangDetect, whatlang и др. Друг експеримент на Jennifer Williams and Charlie K. Dagli[3] за разпознаване на език и диалект измежду сходни такива използва анотиране, чрез първоначална класификация на съобщенията според геолокацията и дава по-прецизни резултати спрямо алгоритъма на Twitter.

При апробирането на алгоритъма на Twitter за разпознаване на естествени езици, вкл. и български, са използвани сборниците на JRC-Acquis[4], съдържащи българското законодателство от 1958 до 2006-та година в електронен вариант. В следствие на това може да изградим две хипотези. Първо, дори и да използваме филтър за език при извличане на съобщения от Twitter, има вероятност алгоритъмът на Twitter да разпознае грешно даден език и да го изпрати към онлайн емисията. Очакваме това да са туитове на други езици, които използват кирилица. Второ, алгоритъмът трябва да дава добри резултати при туитове с политически или правен контекст, т.е. може да се очаква значителна част от съобщенията да са на такава тематика.

С оглед на несъвършеността на алгоритмите за разпознаване на естествени езици в микротекст е необходимо след колекциониране на туитовете да се направи допълнително пречистване. След извършване на предварителна обработка за извличането топ тенденциите в дискусиите на български език предлагаме клъстеризиране на текстовата част на туитовете. Сходни разработки в областта са свързани с извличане на тематични дискусии по време на световното първенство по футбол (Godfrey et al. 2014); разпознаване на възникващи световни събития (Ifrim et al, 2014); разпознаване на спам съобщения в Twitter (Miller et al, 2014); клъстеризация на новини (Rosa et al, 2011) и др.

2. Провеждане на експеримент

За достъп до съобщенията в Twitter, платформата предлага на разработчиците 4 приложни интерфейса:

- Ads API – за създаване и управление на рекламни кампании;
- Streaming API – за филтриране на съобщения в реално време, дава възможност за абониране към публичните емисии от данни;

[2] Lui M, T. Baldwin, Accurate Language Identification of Twitter Messages. Proceedings of the 5th workshop on language analysis for social media 2014 p. 17-25

[3] Williams J, C. Dagli, Twitter Language Identification Of Similar Languages And Dialects Without Ground Truth. VarDial. 2017, p. 73.

[4] The Acquis Communautaire, <https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis>, 19.10.2017

- Search API – за търсене в архиви от състояли се в отминал период туитове;
- Direct Message API – за работа със системата за лични съобщения между потребителите;

С помощта на Streaming API се изтеглят съобщения от Twitter в момента на тяхното публикуване. Този интерфейс дава между 1% и 4% от реалната публична емисия, като за корпоративни клиент се предлага и платен вариант, с който се гарантира 100% интегритет на извлечените данни. Ограничението е наложено, поради големия обем и скорост на публикуване на нови съобщения в платформата. За целите на експеримента в продължение на десет дни (от 01.10.2017 до 10.10.2017) между 09:00 и 21:00 часа са извличани съобщения, съдържащи български текст, според алгоритъма на Twitter. Като допълнително условие за таргетиране на съобщения са включени най-често срещаните служебни думи от българския език – “и”, “е”, “с”, “в”, “на”, “от”. Туитовете извличат в JSON формат, като туитовете за един ден се съхраняват в един и същи файл. След изтегляне на съобщението сме избрали да запишем определени атрибути в база от данни – номер, тяло на съобщението, потребителско име на автора, локация, дата, брой споделяния, брой харесвания, брой коментари, брой цитирания. Тук е важно да уточним, че при създаване на туит броя харесвания, коментари и т.н. е 0, тъй като веднага след публикуването му се то изпраща на публичната емисия. Ако в следващ момент постъпи коментар или споделяне на това съобщение, в емисията влизат и актуалните стойности на тези атрибути.

Първичната обработка включва премахването на съобщения, които могат да бъдат идентифицирани като несъдържащи български език. Това са съобщения, идващи от Русия, Македония, Сърбия и други държави, използващи кирилица и букви, които не присъстват в българския език като напр. руското “ы” или македонското “ја”. Следващата стъпка включва премахването специфични текстови низове или поредици от символи, характерни за неформалната реч в социалните мрежи. С помощта на регулярни изрази се премахват последователно всички:

- споменавания на фразата за споделяне на съобщение “RT” (RT .*?:);
- споменавания на потребители (\s*@(.+?)\s);
- линкове (http(.*)\b);
- символи, които не са част от латиницата или кирилицата ([\p{Z}]{2, }/u);
- често срещани емотиконки;
- числа ([0-9]+);
- пунктуация отговаряща на POSIX ([:punct:]);
- допълнителна пунктуация („|“|–|”|…|«|»|—);
- думи по-кратки от 3 символа (ако текстът е в Unicode формат се удвоява дължината, т.е. по-кратки от 6 символа);
- най-често срещани служебни думи в българския език – “през”, “като”, “след”, “това”, “този”, “също”, “само”, “нещо”, “може”

Първичната обработка завършва с съхранението на съобщенията в JSON файл, където на всеки отделен ред е записан по 1 туит. Документите се клъстеризират веднъж индивидуално и след това заедно. Използваният алгоритъм на клъстеризация е k-means++ с 5 клъстерни центъра и максимум 100 итерации. Текстовите съобщения се превръщат във вектори, като посоката на вектора се определя с индекса td-idf (Term Frequency – Inverse Document Frequency или честота на термина – обратна честота на документа) за отделните думи в туита. След извършване на клъстеризацията се извеждат 5 ключови думи от всеки клъстер.

3. Резултати

В Таблица 2 са поместени данни за броя извлечени съобщения и филтрирани съобщения, както и процентно отношение за това каква част от всички туитове са на руски, македонски или сръбски.

Таблица 2. Брой извлечени и филтрирани съобщения

Дата	01.10	02.10	03.10	04.10	05.10	06.10	07.10	08.10	09.10	10.10
Брой извлечени съобщения	8819	15276	9392	14121	15232	11179	11897	11249	16028	14290
Брой след филтриране по локация	7358	13260	8295	12361	13285	9818	10365	9585	14169	12549
% отстранени по локация	16.57%	13.20%	11.68%	12.46%	12.78%	12.17%	12.88%	14.79%	11.60%	12.18%

Заключение / Conclusion

Около 10-12 реда!!! Определянето на интервала от време за презареждане се осъществява чрез съобразяване с две взаимно противоречащи си условия: висока степен на актуалност на данните (минимална латентност) при ниско натоварване на уеб сървъра. В идеалния случай интервалът на презареждане трябва да съвпада с интервала на възникване на събития при сървъра. Определянето на интервала от време за презареждане се осъществява чрез съобразяване с две взаимно противоречащи си условия: висока степен на актуалност на данните (минимална латентност) при ниско натоварване на уеб сървъра. По-дългият интервал може да доведе до пропускане на точния момент на възникване на някое събитие при сървъра, а по-късият интервал води до по-голям трафик и натоварване на уеб сървъра. В идеалния случай интервалът на презареждане трябва да съвпада с интервала на възникване на събития при сървъра. По-дългият интервал може да доведе до пропускане на точния момент на възникване на някое събитие при сървъра, а по-късият интервал води до по-голям трафик и натоварване на уеб сървъра.

Използвана литература

1. Ellen J. All about Microtext-A Working Definition and a Survey of Current Microtext Research within Artificial Intelligence and Natural Language Processing. ICAART (1). 2011, p. 329-36.
2. Godfrey D, C. Johns, C. Sadek, Interpreting Clusters of World Cup Tweets.2014
3. Ifrim G, B. Shi, I. Brigadir, Event detection in twitter using aggressive filtering and hierarchical tweet clustering. InSecond Workshop on Social News on the Web (SNOW), Seoul, Korea, 8 April 2014 2014 Apr 8. ACM.
4. Lui M, T. Baldwin, Accurate language identification of twitter messages. InProceedings of the 5th workshop on language analysis for social media (LASM)@ EACL 2014 Apr 26 (pp. 17-25).
5. Rosa K., R. Shah, B. Lin, A. Gershman, R. Frederking, Topical clustering of tweets. Proceedings of the ACM SIGIR: SWSM. 2011 Jul.
6. Williams J, C. Dagli, Twitter Language Identification Of Similar Languages And Dialects Without Ground Truth. VarDial 2017, p. 73.
7. Xue Z., D. Yin, B. Davison, Normalizing Microtext. Analyzing Microtext. 2011, 8, p. 5.
8. Yin J. Clustering Microtext Streams for Event Identification. InIJCNLP. 2013, p. 719-725

За контакти
ас. Борис Банков
Икономически университет – Варна
boris.bankov@ue-varna.bg