Извор: https://www.aclweb.org/anthology/L18-1506.pdf

ТЕРМИНИ:

The Recall-Oriented Understudy for Gisting Evaluation (Rouge)-score -> скаларна вредност во ранг од [0,1]. ја прикажува сличноста помеѓу еден и еден или повеќе токенизирани документи. Го оценува квалитетот на моделите за сумирање.
Сентиментален - нешто што одма ми ги разгорело луѓето. Пример, ако напишеш твит, Donald Trump се самоубил, тоа е сентиментален твит.

ВОВЕД

Во воведот се објаснува за претходни истражувања во врска со текст сумаризацијата. Проблемот настанувал поради мали податочни множества за проценка кој е подобар модел за сумаризација. Пример за твитовите, авторите морале сами да создаваат податочни множества. Иако имало множество од твитови со преку 230 000, ипак имало проблем со референците до нив што влијаеле врз проценките на моделот.

Овој научен труд уствари претставува текст сумаризација, прикажувајќи ново множество на твитови кои обработуваат 6 теми од социјален аспект. За да покажат дека овој dataset е уствари потенцијално употреблив, користат хибриден TF-IDF за извлекување на нови твитови, за да се докаже дека има компетитивен ROUGE скор.

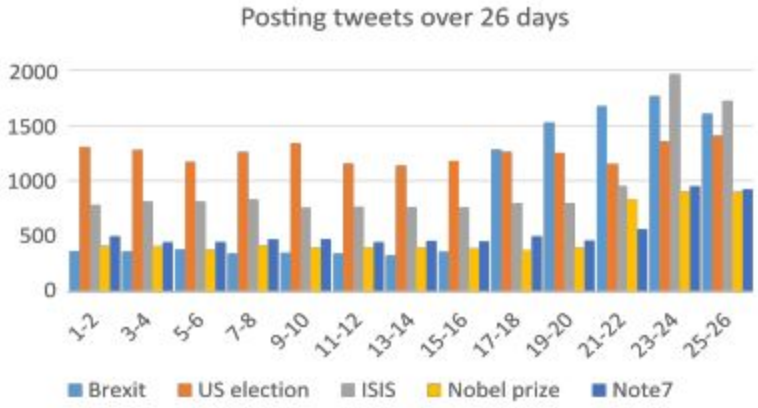
СУМАРИЗАЦИЈА:

Се дели на два чекора: data creation и моделот вклучувајќи и tweet scoring, selection

Data Creation

Значи за да се направи под. М-во потребно е да листа од области кои треба задоволуваат некои услови. Како на пример, да бидат преземени од различни извори, не се постари од 30 дена и мора да бидат

некои популарни (импресивни) да речеме..Потоа за секоја област (topic) се доделува листа со клучни зборови, поединечно.



Event	#tweets	#hashtags	Keywords
Brexit	10,978	9,705	brexit, #brexitshambles, #Brexiters, #BrexitCentral, England Europe exit.
US election	17,714	8,566	Donald Trump, Trump, Hilary Clinton, Clinton #debate, election.
ISIS	13,047	9488	ISIS, IS Syria, IS Mosul, IS Iraq, ISIS Aleppo, ISIS US, ISIS Rusia.
SS Note 7	7,362	7,465	Galaxy Note 7, #note7, #GalaxyNote7, thegalaxynote7, #SamsungGalaxyNote7.
Nobel prize	6,812	2,780	Nobel prize 2016, Nobel peace, Nobel chemistry, Nobel economy, Nobel physics.
SpaceX	4,982	2,417	Facebook SpaceX, SpaceX Explosion, Falcon 9 exploded, Falcon 9 explosion.

Ова е распределба на твитови. Покажува статистика за користење на клучните зборови.
МОЖЕ ДА БИДАТ ИСКОРИСТЕНИ ЗА ДА СЕ ПРОЦЕНИ КВАЛИТЕТОТ НА ТВИТОВИТЕ, МЕРЕЈЌИ КОЛКУ Е БИТНА ИНФОРМАЦИЈАТА КОЈА Е ГЕНЕРИРАНА ОД КОРИСНИЦИТЕ НА ТВИТОВИТЕ.

Податочна сигментација

Собраните твитови се делат во кластери. Идејата е таква бидејќи дури и бројот на твитови на ден да е мал, директно извлекување на подмножество од овие твитови можеби ќе ги елиминира останатите кои се битни. Значи… со кластерирањето, целта е да твитовите бидат што порепрезентивни.
Е сега за да биде се тоа во реално време од аспект на твит сумаризацијата, се користи алгоритам **Affinity Propagation , алгоритам за кластерирање.**
AP алгоритамот, го идентификува подмножеството од data points како темплејтови и формира кластери така што ги доделува останатите data points во еден од тие темплејтови (exemplars). После кластерирањето и елиминирање на тие денови со малку твитови, се формирале 6 множества

кои кореспондираат на 6те настани во тие 26 дена и секој ден вклучува м-во од кластери, кои на некој начин се подобласти.

Standard reference creation

Ова се користи за еволуација на моделот. Се користат 3 методи
1.Luhn кој е хевристички метод за извлекување на тие податоци.
2. LexRank е графички-базиран модел, кој гради граф на сличност и ги селектира оние кои се засноваат на нивниот сопствен вектор (ова е од линеарна алгебра и сега не е толку битно да го знаеш ти).
3. DSDR-non се базира на ненегативна линеарна податочна реконструкција.

Овие извлечени твитови од овие три методи формираат множества кои се кандидати.

Method	ROUGE-1	ROUGE-2	ROUGE-SU
Luhn	0.556	0.502	0.250
LexRank	0.581	0.516	0.273
DSDR-non	0.419	0.310	0.145

Според ова LexRank е најдобар метод, бидејќи има најголема ROUGE вредност

Обзервација на податоците:

Луѓето последователно користеле емотивни зборови и хаштази во нивните твитови. Емотивните зборови ги прикажувале нивните интереси и хаштазите ги користеле како назначување за кој настан зборува.

Поларитетот е уствари позитивно, негативно, неутрално мислење.

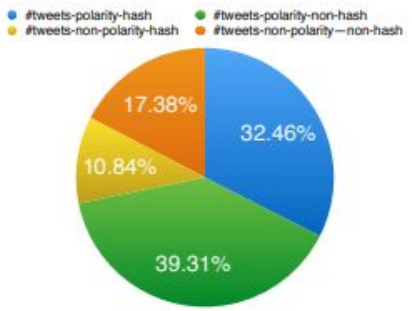


Figure 2: Hashtag and polarity observation on six datasets.

МОДЕЛ

Tweet scoring

$$TF\_IDF = tf_{i,j} * \log_2 \frac{N}{df_j} \qquad (1)$$

$tf_{i,j}$ - фреквенцијата на терминот T_j во документот D_i, каде N е вкупен број на документи и df_j е бројот на документи кој го содржи терминот T_j

Равенката претсавува мерка која пресметува колку е релевантен еден збор во документ во однос на една колекција од документи.
Понатаму, во трудот се објаснува за недостатоците и предностите од аспект на извлекувањето на твитовите. Поради тие недостатоци, се усвојува нов метод , односно хибриден TF-IDF. Разликата е во тоа што, сите твитови се сметаат како еден единствен документ при пресметување на TF, а секој твит како посебен документ при пресметување на IDF.

$$h_{TFIDF}(t) = \frac{\sum_{i=0}^{\#WordsInTweet} W(w_i)}{nf(t)} \quad (2)$$

$$W(w_i) = tf(w_i) * \log_2(idf(w_i)) \quad (3)$$

$$tf(w_i) = \frac{\#OccurencesOfWordInAllTweets}{\#WordsInAllTweets} \quad (4)$$

$$idf(w_i) = \frac{\#Tweets}{\#TweetsInWhichWordOccurs} \quad (5)$$

$$nf(t) = \#WordsInTweet \quad (6)$$

w_i е i-тиот термин во твитот t, W() враќа тежина на терминот, tf() враќа TF скор, соодветно idf() и nf() е некаков фактор на нормализација на твитот, бидејќи се долги твитовите.

Tweet selection

Се одбираат top m рангирани твитови по најголемите скорови кои се селектирани како сумаризација за секој кластер(документ)

Експеримент и резултати

Се споредуваат и други методи

Method	ROUGE-1	ROUGE-2	ROUGE-SU
KL	0.394	0.263	0.146
LSA	0.462	0.368	0.175
Sumbasic	0.444	0.298	0.174
TextRank	**0.495**	**0.418**	**0.213**
Retweet	0.384	0.264	0.129
DSDR-lin	0.460	0.351	0.183
h-TFIDF	*0.482*	*0.384*	*0.199*

ROUGE scores with hashtags We also evaluated all the methods by using hashtags. The intuition is that tweets usually include hashtags, which show important information regarding user's interests. To do that, we extracted all hashtags of each cluster to form its artificial references.

Table 4: The average ROUGE scores over six datasets.

Method	ROUGE-1	ROUGE-2	ROUGE-SU
KL	0.122	0.033	0.015
LSA	0.111	*0.034*	*0.017*
Sumbasic	**0.137**	*0.034*	*0.018*
TextRank	0.100	0.030	0.011
Retweet	0.101	0.024	0.007
DSDR-lin	0.117	0.033	0.011
DSDR-non	*0.123*	**0.036**	0.010
Luhn	0.105	0.032	0.011
LexRank	0.118	0.033	0.013
h-TFIDF	0.113	0.031	0.010

Обзервација на поларитетот кај твитовите

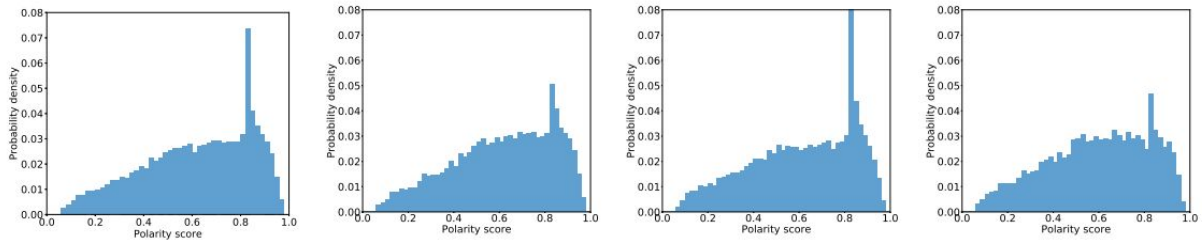


Figure 3: Polarity distribution of original and extracted tweets from three models.

Се тренира модел на класификатор за да види колку и каков е поларитетот на твитовите. Значи се зима м-вото од податоци, се тренира CNN, потоа се предвидува, тој скор што ќе се добие усвари е конвертиран во ранг од [0,1] и тоа е интизитетот на поларитет, каде твитовите со високи скорови се всушност несентименталните твитови, а оние кои се до приближно 0 се сентиментални.

Различни методи тестирани според интезитетот поларитетот на твитот заедно со нивниот закон на распределба.

ИЗВОР: https://arxiv.org/pdf/1511.08417.pdf

АБСТРАКЦИЈА

Слично како минатиот труд во однос на тоа како се дели м-во на податоци, но овде наместо хибриден TF-IDF, се користи Integer Linear Programming (ILP).
Секој хаштаг на твит претставува кластер од документи за различни topics.

ВОВЕД

Значи проблемот е што нема доволно сумаризации направени од човек кои биле користени за тренинг. Во глобала се прави анализа(по нашки муабет) на твитовите. Се претставува TGSum што претставува мулти документирана сумаризација на податочно м-во од твитови. Со користење на ILP се открива м-во од реченици со кои ќе се добие горна граница на ROUGE.

Dataset	Cluster #	Doc. #	Sent. #	Ref. #
DUC 01	30	309	10639	60
DUC 02	59	567	15188	116
DUC 04	50	500	13129	200
TGSum	204	1114	33968	4658(tweets)

Table 2: Statistics of the summarization datasets.

TGSum Конструкција

4 чекори:

- URL Acquisition
- Data Collection
    - Освен извлекување на твитови треба и предпроцесирање на иститите:
        - Одбивање на ретвитови
        - Бришење на токени кои не се на Англиски во твитот
        - Бришење на твитови кои содржат помалку од 5 токена
        - Спојување на идентични твитови.
- Cluster Formation
- Reference Generation

Analysis of ROUGE Measurement
Extension to Multiple ROUGE Variants
Математиката е позади оваа мерка, за различни варијатни на ROUGE, доволно да знаеме што значи неговиот исход.

ЕКСПЕРИМЕНТ

4 типа на различни извори за твитови:
-Извлекување -> твитот е директно извлечен од оргинален текст
-Компресија -> Иста секвенца од зборови во твит може да биде најден во документ
-Абстракција -> Преку 80% од зборовите во твитот може да бидат најдени во новостите (документ) и лоциран во повеќе од една реченица.
-Останато-> тоа во глобала се коментари

Type	Tweet	Source
Extraction	Police have released a sketch of the main suspect	**Police have released a sketch of the main suspect**, a man in a yellow T-shirt who was filmed by security cameras leaving a backpack at the shrine.
Compression	Suspect in Bangkok bombing is "an unnamed male foreigner," according to an arrest warrant issued by a Thai court.	The chief **suspect in** the deadly **bombing** of **Bangkok**'s popular Erawan Shrine **is "an unnamed male foreigner," according to an arrest warrant issued** Wednesday **by a Thai court.**
Abstraction	Taxi driver who thinks he picked up Bangkok bombing suspect says man was calm, spoke unfamiliar language on a phone.	A Thai motorbike **taxi driver who believes he picked up the suspect** shortly after the blast also said he did not seem to be Thai. ... who **spoke an unfamiliar language on** his cell **phone** during the short ride ... ... he still appeared very **calm**...
Other(Comment)	Hmm, this face looks a bit familiar...	NULL

Table 3: Examples of different linked tweet types.

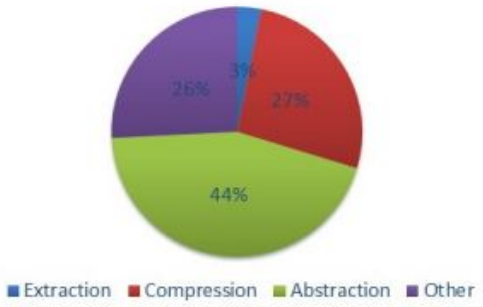


Figure 1: Proportions of linked tweet types.

Test set	Training set	ROUGE-1	ROUGE-2
01	TWT_REF	29.23	5.45
	TGSum	29.40	5.73
	DUC	29.78	6.01
	DUC+TWT_REF	30.13	6.08
	DUC+TGSum	**30.32**	**6.26**
02	TWT_REF	31.10	6.34
	TGSum	31.71	6.73
	DUC	31.56	6.78
	DUC+TWT_REF	31.74	6.80
	DUC+TGSum	**32.15**	**6.89**
04	TWT_REF	35.73	8.83
	TGSum	35.97	9.16
	DUC	36.18	9.34
	DUC+TWT_REF	36.19	9.23
	DUC+TGSum	**36.64**	**9.51**

Table 6: Summarization performance with different training data.