Combined analysis of news and Twitter messages

Mian Du Jussi Kangasharju Ossi Karkulahti Lidia Pivovarova Roman Yangarber
University of Helsinki, Department of Computer Science

Abstract

While it is widely recognized that streams of social media messages contain valuable information, such as important trends in the users' interest in consumer products and markets, uncovering such trends is problematic, due to the extreme volumes of messages in such media. In the case Twitter messages, following the interest in relation to all known products all the time is technically infeasible. IE narrows topics to search. In this paper, we present experiments on using deeper NLP-based processing of product-related events mentioned in news streams to restrict the volume of tweets that need to be considered, to make the problem more tractable. Our goal is to analyze whether such a combined approach can help reveal correlations and how they may be captured.

1 Introduction

Twitter is a social networking and a microblogging service, that currently has more than 500 million users of which 200 million are using the service regularly. Many commercial organizations e.g. companies, newspapers and TV stations, as well as public entities, publish and promote their content through Twitter. According to the company itself, 60% of its users "access the service through mobile devices." On Twitter the relationships are by default directed, that is, user A can follow user B's posts without B following A. The posts on Twitter are referred to as tweets and at the moment of this writing there are more than 500 million tweets created daily. A tweet is limited to 140 characters of text, a legacy from the time when the system was envisioned to operate via SMS messages.

We will argue that our practice is not applicable to the Twitter service exclusively, however we

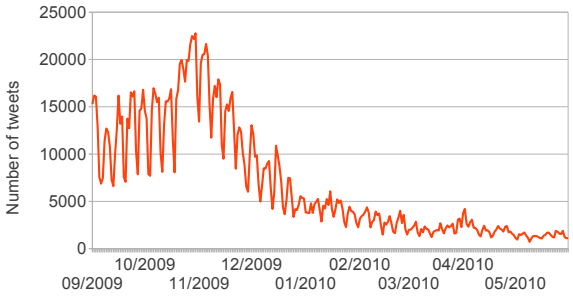


Figure 1: H1N1 on Twitter

elected to survey Twitter for a number of reasons. It has a huge number of users and is used worldwide. Because tweets are limited in length, the amount of data to be collected is kept manageable and it also helps maintain the analysis process simple. However, the most important factor for us was its openness. By default all tweets are public and the service offers a relatively functional and free API for gathering data.

Our earlier investigations demonstrate that Twitter users do react with higher volumes of posts to topical, news-worthy events. For instance, consider Figure 1, which plots the number of posts that contain keywords related to the 2009-2010 outbreak of H1N1 virus (swine flu). The curve matches almost perfectly with the peak of the outbreak and declines as the epidemic decayed. In this paper we will show that the topicality can be extended to business events, such as new product releases, and some releases indeed generate a large number of posts.

Until recently, a large part of research on social media has focused on analyzing and examining networks and graphs that emerge among users, references and links, and measuring patterns in creation and consumption of content. At present, more attention is being devoted to analyzing the vast volume of messages in the social

Proceedings of the Joint Workshop on NLP&LOD and SWAIE, pages 41–48,
Hissar, Bulgaria, 12 September 2013.

media in terms of the *content* of the messages itself. Researchers in academia and industry are eager to mine the content for information that is not available from other sources, or before it becomes available from other sources (for example, see (Becker et al., 2012; Becker et al., 2011), and other works of the authors).

However, our work is not aimed at event discovery in Twitter. Instead, we try to discover how events, which we find in other sources—e.g., in traditional media—are presented on Twitter. We assume that it is worthy to know not only what kind of events can be found in Twitter but also events that are not present in tweets. For example, continuing the previous example, we can note that apart from flu there are many other diseases that can be less represented or completely absent from tweets.

From the point of view of natural language processing (NLP), the immediate problem that arises is that the linguistic register and language usage that is typical for social media content—such as web logs, and especially the ultra-short messages, such as those on Twitter—is very different from the register and usage in "traditional," well-studied sources of on-line textual information, such as news feeds. Therefore, it has been observed that new approaches are needed if we are to succeed raising the quality of analysis of the content of social media messages to useful levels. This territory remains largely uncharted, though the need is quite urgent, since a better understanding of the content will enable developments in areas such as market research and advertisement, and will also help improve the social media services themselves.

In this paper we examine how companies and products mentioned in the news are portrayed in message streams on the Twitter social networking service; in particular, we focus on media events related to the announcement or release of new products by companies. Our main research questions are: do interesting correlations exist between reports of a product release in the news and the volume of posts discussing the product on Twitter? Are some types of products more likely to generate more posts than others? Do different types of products trigger the generation of different types of messages?

One serious problem when conducting social media research is managing the data collection, and assuring that the system does not become overwhelmed with an enormous volume of data. In this paper we present a hybrid approach, where we first apply Information Extraction (IE) to messages found in news streams to narrow down scope of potentially relevant data that we will subsequently collect from Twitter. The volume of news is orders of magnitude smaller and more manageable than the volume of Twitter. In particular, extracting company and product names mentioned in the news will yield keywords that will match hot topics on Twitter. Although we may miss some important events on Twitter using this procedure, we reason that it is more tractable than continually keeping track of a large list of companies and products. An equally important factor is the fact that keeping lists of companies and products is not only impractical, but it is also insufficient, since new companies and novel products are introduced to the markets every day.

Our contributions and results include:

- we demonstrate how deeper NLP analysis can be used to help narrow down scope of messages to be retrieved from social-media message streams;
- we observe interesting correlations between events that are found in the two sources;
- we present some details about the content of tweets that correspond to news-worthy events: e.g., proportions retweeted messages and links, showing that sharing links is common when discussing certain products.

The remainder of the paper is structured as follows. Section 2 discusses related work. Section 3 describes the event extraction process, and covers the details of the data collection from Twitter. We discuss our results in Section 4, and Section 5 presents our conclusions and an outline of future work.

2 Related work

Research on social media, and on Twitter in particular, has been attracting increasing attention. It is a crucial source of information about public moods and opinions, for example, on topics of public concern such as political changes and elections (Diakopoulos and Shamma, 2010), or revolutions (Lotan et al., 2011). Twitter also can be

useful for monitoring of natural disasters and epidemics of infectious disease (Lamb et al., 2013). At the same time, Twitter is a problematic source since traditional NLP methods for information extraction, opinion mining, etc., are not directly applicable to very short texts, or texts using communication styles peculiar to social media (Timonen et al., 2011).

Similar work to ours was reported by Tanev et al. (2012), who first used a fact extraction system to find events related to social unrest and cross-border criminal activity, and then tried to find additional information by using Twitter feeds. Becker et al. (2011) trained a classifier to distinguish tweets that relate to real-word events from tweets that do not. They demonstrate that event-related tweets are quite rare; the majority of tweets do not contain events.

Kwak et al. (2010) compared topics that attract major attention on Twitter with coverage in other sources, namely, Google Trends and CNN headlines. They have found that Twitter can be a source of breaking news as well. Zhao et al. (2011) used topic modelling to compare Twitter with the New York Times news site. They found Business as being among the top-10 topics on Twitter; however, business-related tweets rarely express opinions.

Krüger et al. (2012) manually analyzed 500 random tweets related to Adidas, and came to the conclusion that the company uses Twitter to promote their brand. Jansen et al. (2009) manually prepared a list of companies and brands belonging to different Business sectors, and then collected tweets related to these companies and brands. They demonstrate that approximately 20% of tweets contain mentions of companies or brands, which means that Twitter is an important marketing medium; however, only 20% of the tweets that mention companies and brands express a sentiment about them.

3 Data Collection

We use PULS[1] to extract events from text. PULS is a framework for discovering, aggregating, vizualization and verification of events in various domains, including Epidemics Surveillance, Cross-Border Security and Business.

In Business scenario events typically include merges and acquisitions, investments, layoffs, nominations, etc. In this paper we focus on "New Product" events, i.e., when a company launches a new product or service on the market. Figure 2 presents an example of a piece of text from a news article and an event structure extracted from this text. A product event describes a company name, a product name, a location, a date, and the industry sector to which the event is related. These slots are filled by a combination of rule-based and supervised-learning approaches (Grishman et al., 2002; Yangarber, 2003; Huttunen et al., 2013).

On Friday, **Nokia** unveiled the **Lumia 928** for the **U.S.** market, priced at $99 after a rebate and a two-year deal with Verizon Wireless.

The Lumia 928 is the latest version in Nokia's range of **smartphones** using Windows Phone software, with its metal body setting it apart from earlier models.

COUNTRY:	US
DATE:	2013.05.10
COMPANY:	NOKIA
PRODUCT NAME:	Lumia 928
PRODUCT DESCRIPTION:	smartphone
SECTOR:	Telecommunications

Figure 2: A news text and a "New Product" event, extracted from this document by IE system.

For identifying the industry sectors to which the events relate, we use a classification system, currently containing 40 broad sectors, e.g., "Electronics," "Food," or "Transport." This classification system is similar to existing classification standards, such as the Global Industry Classification System (GICS),[2], or the Industry Classification Benchmark (ICB, http://www.icbenchmark.com/), with some simplifying modifications. The sector is assigned to the event using a Naive-Bayes classifier, which is trained on a manually-labeled set of news articles, approximately 200 for each sector, that we collected over several years.

We use the new-product events extracted by PULS to construct special queries to the Twitter API. One query contains a company name and a product name, which are the slots of a product event (see Figure 2). Every day we extract about 50 product events from news articles, and generate 50 corresponding queries to the Twitter API. We then use the Twitter API and collect all tweets that include both the company and the product name. Below one can see an example tweet containing the company name *Audi* and the product name *A3*:

[1]The Pattern Understanding and learning System: http://puls.cs.helsinki.fi

[2]http://www.msci.com/products/indices/sector/gics/

Time	Events	Tweets
Nov 2012–May 2013	1764	3,842,148

Table 1: Dataset description

```
The new A3 from Audi looks great!
```

The Twitter API has some restrictions. While conducting our survey[3], we could make 150 requests per hour, asking for 100 tweets per request, yielding a maximum of 15,000 tweets per hour. We had at our disposal the University of Helsinki cluster consisting of approximately machines, giving us the theoretical possibility to collect up to 3,000,000 tweets per hour.

While the company and product names are used as keywords in the Twitter query, other slots of the event are used for analyzing the results of the query. These slots, which include the industry sector, the country, the product description, and the date of the report, are used to label the tweets returned by the query. For example, we extract an event as in Figure 2 and get 2,000 tweets which contain both "Nokia" and "Lumia 928". Since the event is related to the industry sector "Telecommunications", we consider these 2,000 tweets are also related to "Telecommunications". Thus, we can group the returned tweets by industry sectors, country, etc., and analyze the flow of information.

The Twitter API lets us fetch tweets from seven previous days, and we kept collecting the tweets for each keyword for at least 3 days after its mention in the news. Thus, every keyword query has a time-line of roughly ten days around the news date.

The dataset is summarized in Table 1. We started the survey in November 2012 and the results presented in this paper include data collected through May 2013. In total, there are 1764 different events and in total close to 4 million tweets. In the final section of this paper we will discuss how we plan to improve the data collection in the future.

4 Experiments and results

4.1 Tweet statistics overview

First we present an overview of the tweet statistics. Table 2 summarizes the statistics, grouping the events based on the number of tweets they generated. The table also lists the total number of tweets, the percent of tweets that contain at least one hyper-link URL, and the percent of "retweets". A retweet is somewhat analogous to forwarding of an email. A retweets starts with "RT" abbreviation, making it easily distinguishable. Note that retweet can contain additional text compared to the original tweet, e.g., the retweeting user's personal opinion. The last column on the table represents the fraction of unique tweets; to count this number we subtracted from the total amount of tweets the number of tweets which were exactly identical. We pruned away the shortened link URLs from the tweet text when we calculated the uniqueness percentage, since the same URL can be shortened differently.

Number of tweets	Number of events	Links %	Retweets %	Unique tweets %
10k+	33	82	22	52
1k-10k	68	78	23	53
100-1k	109	79	24	61
10-100	258	84	18	73
1-10	249	85	12	85

Table 2: Overall statistics: number of tweets, links and retweets per event

As can be seen from Table 2 there were 33 product events that generated more than 10,000 tweets. Strikingly, 82 percent of the tweets had a link. We checked a random sample through a subset of the tweets, and it seems that the single most common reason for the high number of links is that many websites today have a "share on Twitter" button, which allows a user to share a Web article with his/her followers by posting it on the user's Twitter page. The resulting tweet will have the article's original title, a generic description of the article (such as the one used in a RSS feed), and a link to the actual article. This can also be seen on the last column in Table 2, since the resulting tweets are always identical.

It is interesting to observe that the tweet uniqueness drops as the number of tweets increases. This would seem to indicate that the likelihood that an article is shared increases with the number of times it has already been shared. The same seems to hold for retweets as well. This corresponds to the observations found in literature: it was shown, (Kwak et al., 2010), that if a particular tweet has been retweeted once, it is likely that it will be retweeted again. Similarly, tweets that contain a URL are more likely to be retweeted (Suh et al., 2010). However, tweets related to business are rarely retweeted (Zhao et al., 2011).

[3]The access conditions have been recently changed

COMPANY	# events	max # tweets	total # tweets
Facebook	13	444188	1931445
Microsoft	18	440831	447104
Google	24	410986	877842
Nokia	8	52955	60655
Nintendo	2	46611	75275
Apple	8	19619	42243
Lamborghini	1	21951	21951
Adobe	3	16230	17801
Lego	2	15371	26001
Audi	9	13373	13829
Netflix	2	9880	14249
Casio	1	8970	8970
Amazon	5	8678	10079
Huawei	5	8559	8906
Sony	12	8081	12459
T-Mobile	2	7884	9043
Adidas	13	6487	9171
Acer	1	6099	8592
Volkswagen	2	4454	4454
Subaru	1	4397	4397
Macklemore	1	4301	4301
Zynga	2	4166	4170
Starbucks	1	3993	3993
Lenovo	2	3129	3129
Land Rover	3	2951	4619
Seat	1	2641	2641
Walmart	1	2575	2575
Samsung Electronics	24	2566	4578
Chevrolet	2	2517	2558
Coca-Cola	23	2432	5891
Deezer	1	2107	2107
Tesla Motors	1	2082	2082
Macef	1	2073	2073
Telefonica	6	2065	2090
Orange	7	1958	2532
H&M	2	1787	1787
Dacia	2	1650	1849
Intel	2	1649	1649
Dell	2	1074	2450
Lacoste	2	799	821

Table 3: Most frequently tweeted companies

4.2 What is tweeted most frequently

The total number of distinct companies present in our data set is 1,140. The majority of these companies occur in one event only; for 50% of the companies less than 10 tweets have been returned. The list of most frequently tweeted companies is shown in Table 3. We show the number of events for a company in our dataset, the maximum number of tweets for any one event, and the total number of tweets for the company.

It can be seen from the table that only events related to well-known IT giants, (Facebook, Google, Microsoft), produce more than 100,000 tweets. Nokia, which is on the fourth position, produces 8 times fewer tweets than Google.[4]

[4]We have found relatively few tweets related to Samsung Electronics, even though these events are about launching new smartphones and other gadgets, which seem to be very popular in Twitter. We believe that we did not find more tweets because the full name of the company—"Samsung Electronics"—is rarely used in the tweets, which tend to refer to it as "Samsung;" this type of synonymy will be taken into account in future work. The majority of tweets related to Samsung are links to news (see an example in Figure 3); the text of these tweets are mostly identical (Figure 4), which means that people do not type new information but only click the "tweet" button on the news page.

SECTOR	events	max # tweets	total # tweets
Media, Information Services	109	444188	1534300
Telecommunications	122	337776	531920
Information Technology	33	169086	182408
Consumer Goods	41	15371	29440
Drinks	94	3993	10312
Automotive Engineering	66	4454	10098
Transport	36	1714	9570
Cosmetics & Chemicals	113	3480	6194
Food	106	4369	5751
Energy	6	277	374
Finance	45	179	316
Textiles	10	166	290
Health	25	81	239

Table 4: Most frequently tweeted industry sectors.

Other companies in table are telecommunication and automotive companies, food and drink producers, cosmetics and clothing suppliers. By contrast airlines receive little attention, the news about opening new flight routes cause little response on Twitter. For example, the only tweet related to a new flight by Air Baltic between Riga and Olbia was found in a Twitter account which is specialized for the airline's news.

The list of the most frequently tweeted industry sectors is shown in Table 4. Note, that the business sectors are assigned to events, not to a particular company; for example, an event that describes Facebook launched "Home," an operating system for mobile phones, was assigned with the sector "Telecommunications Technologies", while an event that describes that Facebook launched Graph Search was assigned with sector "Media, Information Services".

As can be seen from Table 4, the sectors in our data are distributed approximately according to Zipf's law: the majority of tweets are related to a limited number of sectors, while the majority of sectors trigger little or no response on Twitter. For example, we do not find any tweets related to such sectors as "Construction" or "Minerals & Metals"; the "Agriculture" sector generated only 3 tweets.

Comparing tables 4 and 3 we can observe that there is a dependency between the number of

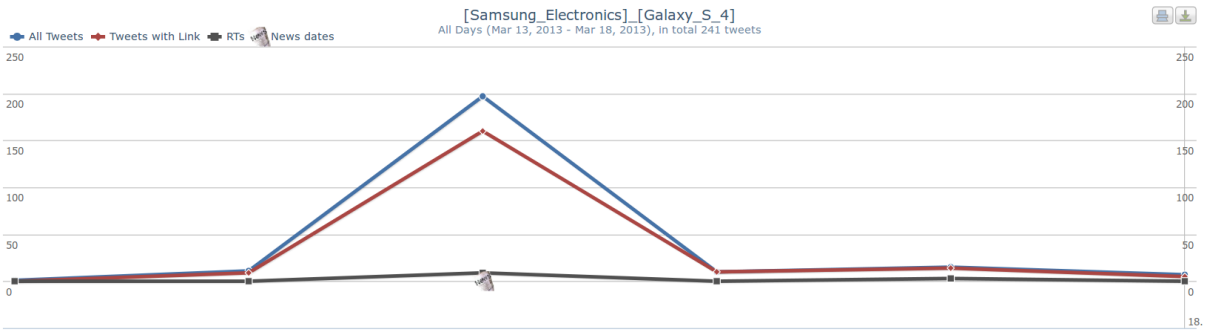


Figure 3: Number of tweets, links and retweets related to an event "Samsung Electronics launched Galaxy S4".

Timestamp (sort)	ID	Tweet	
2013-03-15T17:59:31	312624113040625665	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line s... http://t.co/98XgVNyfpp	campus42
2013-03-15T17:59:11	312624028839981057	Noticias TNO Venezuela - Samsung presenta el GALAXY S 4: Samsung Electronics anunció hoy la cua... http://t.co/5iL3luXEJf #noticias #tno	Grupotno
2013-03-15T17:56:59	312623476605341696	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line s... http://t.co/MzBzNoqzxa	World News
2013-03-15T17:56:57	312623465184235521	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line s... http://t.co/aMFxz7ZmEC	SantinaMelgar
2013-03-15T17:56:54	312623455122124800	#TeamFollowBack How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. re... http://t.co/mLfYFHDsS4 #AutoFollowBack	Vermandita
2013-03-15T17:56:50	312623436721704961	#Tech How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-... http://t.co/FzsjeQjllh	zankrut
2013-03-15T17:56:50	312623435819917312	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed it... http://t.co/nvVT1PXlYW http://t.co/jczlrGv4c1	Riyajain1
2013-03-15T17:56:48	312623429033553921	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line s... http://t.co/nKRRFRzqRn	technewsplace
2013-03-15T17:56:46	312623421320216578	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line s... http://t.co/EgvCvVMNgR	drGalauuu
2013-03-15T17:56:44	312623412621213697	How the Galaxy S 4 stacks up against iPhone, S III: Samsung Electronics Co. revealed its new top-of-the-line… http://t.co/GbSS116OHA	TomFlowers

Figure 4: Tweets related to an event "Samsung Electronics launched Galaxy S4".

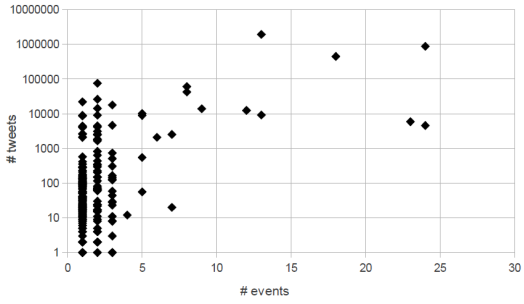


Figure 5: Number of events against total number of tweets for companies.

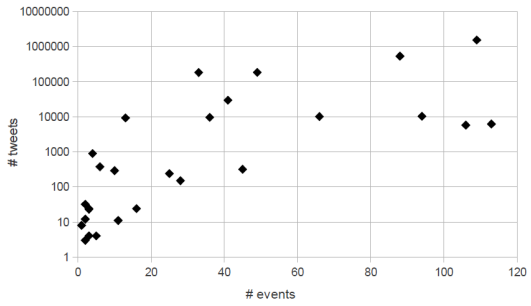
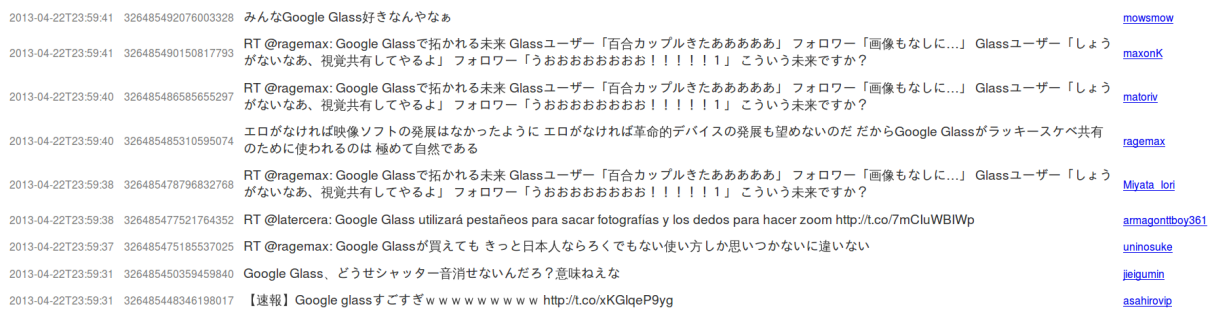


Figure 6: Number of events against total number of tweets for sectors.

events related to a particular sector and the number of tweets related to this sector, whereas there seems to be no such relation between the number of events related to particular company and a number of tweets related to this company. For example, only one event involving Acer appeared during the covered period—a launch of the "Iconia B1" tablet—but it drew more than 6,000 tweets.

The dependencies between the number of events and the number of tweets for companies and sectors are presented in Figures 5 and 6 respectively.

All events were taken from news written in English, but depending on the resulting keywords, the tweets that match the query could be in any language. Since we use the English names for companies and products there is an inherent bias toward countries that use languages with a Latin-based script. However, despite that were able to find many tweets for events that happen in countries that use non-Latin scripts, e.g., Russia or Japan. Two reasons for this may be that the larger companies operates globally, and that Twitter users tend to type company and product names in English even though they tweet in their own languages, see examples in Figure 7.



2013-04-22T23:59:41	326485492076003328	みんなGoogle Glass好きなんやなぁ	mowsmow
2013-04-22T23:59:41	326485490150817793	RT @ragemax: Google Glassで拓かれる未来 Glassユーザー「百合カップルきたあああああ」 フォロワー「画像もなしに…」 Glassユーザー「しょうがないなあ、視覚共有してやるよ」 フォロワー「うおおおおおおお！！！！！1」 こういう未来ですか？	maxonK
2013-04-22T23:59:40	326485486585655297	RT @ragemax: Google Glassで拓かれる未来 Glassユーザー「百合カップルきたあああああ」 フォロワー「画像もなしに…」 Glassユーザー「しょうがないなあ、視覚共有してやるよ」 フォロワー「うおおおおおおお！！！！！1」 こういう未来ですか？	matoriv
2013-04-22T23:59:40	326485485310595074	エロがなければ映像ソフトの発展はなかったように エロがなければ革命的デバイスの発展も望めないのだ だからGoogle Glassがラッキースケベ共有のために使われるのは 極めて自然である	ragemax
2013-04-22T23:59:38	326485478796832768	RT @ragemax: Google Glassで拓かれる未来 Glassユーザー「百合カップルきたあああああ」 フォロワー「画像もなしに…」 Glassユーザー「しょうがないなあ、視覚共有してやるよ」 フォロワー「うおおおおおおお！！！！！1」 こういう未来ですか？	Miyata_lori
2013-04-22T23:59:38	326485477521764352	RT @latercera: Google Glass utilizará pestañeos para sacar fotografías y los dedos para hacer zoom http://t.co/7mCIuWBlWp	armagonttboy361
2013-04-22T23:59:37	326485475185537025	RT @ragemax: Google Glassが買えても きっと日本人ならろくでもない使い方しか思いつかないに違いない	uninosuke
2013-04-22T23:59:31	326485450359459840	Google Glass、どうせシャッター音消せないんだろ？意味ねえな	jieigumin
2013-04-22T23:59:31	326485448346198017	【速報】Google glassすごすぎｗｗｗｗｗｗｗｗｗ http://t.co/xKGlqeP9yg	asahirovip

Figure 7: Tweets related to an event “Google launched Google Glass”.

5 Conclusion and future work

We described an end-to-end framework, which allows us to analyze the influence that business news have on tweets. We have demonstrated that the impact that new-product events have on Twitter depends more on the industry sector than on a particular company. It is clear, however, that the developed framework can be used in more broad applications, at least for more sophisticated data analysis.

Our data, as it was shown before, include the event date and the timestamps for tweets. However, in the current paper this data have been overlooked in the analysis. Thus, the main direction of the further work will focus more on the temporal dimension. We are going to add more metrics, such as the time gap between the product launch and the peak of tweets.

Furthermore, we would like to see whether we could predict the impact created by a product launch based on the history and to find out if there are some models to match that and the lifetime of the tweets. To solve this problem, we plan to modify our data collection process and to monitor a several big companies for a longer time, in order to establish baselines. This will allows us, first, to analyze the exact impact of a product launch on Twitter volume and, second, to measure an impact of corpus narrowing using IE.

Another aspect of the data, which would be interesting to investigate, is location. As have been shown before, the business events include a country slot; however, we cannot assume that corresponding tweets originate from the same country. Thus, we are going to use geolocation techniques, (Dredze et al., 2013; Bergsma et al., 2013), to find the tweets’ countries and to compare them with the countries found in news.

We also plan to improve the query construction algorithm to find more tweets for compound company names, such as “Samsung Electronics.” This cannot be done in a straightforward fashion: “Samsung” may likely refer to “Samsung Electronics”, though “Electronics” may refer to many different entities. Thus we cannot simply search for all substrings of a company name, because such queries will produce to many false hits. We assume that special named entity recognition techniques, which have been recently developed for Twitter (Ritter et al., 2011; Piskorski and Ehrmann, 2013), can be used to solve this problem. To improve coverage it is also possible to use automatic transliteration, which allows to map proper names from Latin to other scripts (Nouri et al., 2013).

We have studied the most and least frequently tweeted companies and industry sectors. In the next phases we will study the most frequently tweeted product types. Since every product found by IE system has a description (as presented in Figure 2), we can group tweets by product type. However, additional work is needed to merge such product types as, for example, “chocolate” and “chocolate candies”. We plan to use a Business concept ontology, which includes the long list of possible product types, to perform this task.

References

Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identification on Twitter. In *International Conference on Weblogs and Social Media*, Barcelona, Spain.

Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. 2012. Identifying content for planned events across social media sites. In *The fifth ACM international conference on Web search and data mining*, pages 533–542, Seattle, Washington.

Shane Bergsma, Mark Dredze, Benjamin Van Durme,

Theresa Wilson, and David Yarowsky. 2013. Broadly improving user classification via communication-based name and location clustering on Twitter. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, Atlanta, Georgia.

Nicholas A. Diakopoulos and David A. Shamma. 2010. Characterizing debate performance via aggregated Twitter sentiment. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1195–1198. ACM.

Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: a Twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002. Event extraction for infectious disease outbreaks. In *Proc. 2nd Human Language Technology Conf. (HLT 2002)*, San Diego, CA.

Silja Huttunen, Arto Vihavainen, Mian Du, and Roman Yangarber. 2013. Predicting relevance of event extraction for the end user. In *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing, pages 163–176. Springer Berlin.

Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Nina Krüger, Stefan Stieglitz, and Tobias Potthoff. 2012. Brand communication in Twitter—a case study on Adidas. In *PACIS 2012 Proceedings*.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of NAACL-HLT*, pages 789–795.

Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and Danah Boyd. 2011. The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, 5:1375–1405.

Javad Nouri, Lidia Pivovarova, and Roman Yangarber. 2013. MDL-based models for transliteration generation. In *SLSP 2013: International Conference on Statistical Language and Speech Processing*, Tarragona, Spain.

Jakub Piskorski and Maud Ehrmann. 2013. On named entity recognition in targeted Twitter streams in Polish. In *The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing : ACL 2013*, pages 84–93, Sofia, Bulgaria.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 177–184. IEEE.

Hristo Tanev, Maud Ehrmann, Jakub Piskorski, and Vanni Zavarella. 2012. Enhancing event descriptions through Twitter mining. In *Sixth International AAAI Conference on Weblogs and Social Media*, pages 587–590.

Mika Timonen, Paula Silvonen, and Melissa Kasari. 2011. Classification of short documents to categorize consumer opinions. In *Proccedings of 7th International Conference on Advanced Data Mining and Applications*, pages 1–14.

Roman Yangarber. 2003. Counter-training in discovery of semantic patterns. In *Proc. ACL-2003*, Sapporo, Japan.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer.