The Pursuit of Newsworthiness on Twitter

Eirik Stavelin
University of Bergen

Abstract

Although journalists follow social media and social media do contain newsworthy insight and knowledge, this type of data is rarely analyzed for journalistic purposes. This paper presents a cluster analysis tool that supports a journalistic inquiry into Twitter messages. By tailoring cluster analysis methods this tool produces subsets of a text collection with similar texts to speed up the quest for interesting user generated content. Results show promise for this kind of computer supported analysis, but also outlines problems both with the methodology and the assumptions that good arguments or novel opinions are enough to be considered newsworthy. Evaluation by professional journalists show that in spite of an acceptable quality of algorithmic sorting, the focus in future applications should put more emphasis personas, and identifying "the usual suspects".

Introduction

Social media has grown to become an arena for anyone to participate in debates and share opinions on any topic. For news outlets, this offers new opportunities to discover stories and issues of debate and public interest. Journalists are among the elite of Twitter users (Kwak et al. 2010; Larsson and Moe 2011) and thus do get insights and stories from this use trough his or hers circle of Twitter contacts. But mostly the data are an ephemeral glimpse of individual tweets. In cases where data from social media can offer new or additional insights, stories or sources for journalist, methods for analyzing data and finding texts of interests are needed. This paper presents experiences with a custom built tool to cluster Twitter texts (tweets) into groups of similar tweets by tailoring clustering techniques for Twitter.

The first obstacle with social media data is the size of the datasets. While Norway is small, and Twitter is not widely used (8% of Norwegian internet user accessed Twitter weekly in the 4th quarter of 2011, (Futsæter 2012)) the datasets exceed what a journalist can be expected to read through in her normal workflow. The second is the unfamiliar mix of participants of debaters and commentators. While a journalist is trained in questioning experts and authorities and finding ordinary people for cases, the cacophony of voices and opinions from everyman on Twitter makes it hard to decide: is it the messages from traditional sources – authorities in various sectors –that are to be examined or the most prominent, popular or clever sayings of 'the public'? Further, methods should be independent of the topics of the analyzed data, and reduce the time needed to get a fair overview over the data.

In order to explore the possibilities to utilize Twitter-data, I built a software tool that puts tweets into chunks of manageable sizes and offers users a navigation interface with simple manipulation functions. The approach is based on the bag of words model in natural language processing, and is at the core an effort to extract a structure in the data that isn't explicitly found as meta-data. The applications take large amounts of tweets and groups those who use similar words into clusters that gets presented to a journalist, so that the journalist can review groups of similar tweets to quicker gain an overview over the material. Functions for further exploration and keeping track of single texts and clusters of texts allows for a rapid analysis of user generated content on Twitter.

This paper was presented at the NIK 2013 conference. For more information see http://www.nik.no/

Related Work

Prior research on Twitter has revealed that actors who are important in the old media landscape also hold key position on Twitter, and Twitter usage follow media events (An et al. 2011; De Longueville, Smith, and Luraschi 2009; Kwak et al. 2010; D. A Shamma, Kennedy, and Churchill 2010). Through the aim of understanding the microblogging phenomenon, different studies have found different user-types (De Choudhury, Diakopoulos, and Naaman 2012; De Longueville, Smith, and Luraschi 2009; Java et al. 2007; Larsson and Moe 2011) and also, that in spite of being a noisy media, a potential for news media to "analyze, interpret and conceptualize a system of collective intelligence, rather than in the established practice of selection and editing of content" (Hermida 2010). In this context tools for journalists has been constructed, both to find sources (Nicholas Diakopoulos, De Choudhury, and Naaman 2012) and to analyze user feedback on media events (N. Diakopoulos et al. 2011). My approach explores the data with intentional blind eye to the authors' status to emphasize the democratic possibility of letting anyone through with their perspective. Ways of detecting events from larger streams of social media data is also developed (H. Becker, Naaman, and Gravano 2011; Weng and Lee 2011) including ways of identifying relevant and useful messages from Twitter (H. Becker, Naaman, and Gravano 2011; Luo, Osborne, and Wang 2012). The studies of politics on micro blogs also reveal that social media usage "shadow" real world events to the degree where election results (Tumasjan et al. 2010) and the change of topics in TV-broadcasted debates (David A. Shamma, Kennedy, and Churchill 2009) can be algorithmically detected. Topics that do not spike, but linger, constantly low-volume ongoing arguments, falls outside of these methods, my approach includes such discourses.

Outside academia relevant projects such as the Knight foundation funded associated press project (The Overview Project (Stray 2012)) also apply clustering techniques to aid analysis of textual data for journalistic enquiry. While the overview project is aimed at optical character recognition (OCR) type corpora, this project aim is events in social media texts.

Much focus has been put on understanding what Twitter is, and who the users of Twitter are and how they do discuss topics that are of interest to the news media (politics, disasters, media events, etc.), but less has been done to apply this to ways of analyzing this to find news stories. I want to explore the possibility to analyze accumulated data for a topic in a way that supports a journalists' need to quickly get an overview over what is discussed.

Design Process

The scenario is simple: a journalist collects tweets for a topic related to her work and end up with too much data and too little time. To read through all texts is unrealistic, so how can she get an overview over what the Twitter-data contains?

The initiation for this project was a presentation of a journalistic project where a team of journalists had analyzed Twitter material concerning the 2011 terrorist attacks in Oslo[1]. The team had spent considerable resources on reading though every single tweet. In such a sensitive case this might be necessary when the data is to be published, but in smaller cases the manual labor of this team can be exchanged with computational means to ensure that this kind of data gets some analytical attention instead of no attention.

[1] The NRK project can be seen at http://nrk.no/terrortwitter

This paper relates to a real world problem and the initial criteria for the prototype was collected in dialogue - by telephone interview - with the team leader from the NRK project and followed up though emails to work out details and adjustments.

The tool should be flexible enough to handle data from all sorts of topics; it should take into account the use of multiple languages used in debates in Norway and offer users a way to assess and store material of particular interest to the journalist. The clusters should represent grouping of the material where similar things are discussed, and reduce the work of looking through one large dataset to look at fewer clusters of similar texts.

While meta-data such as time and location can aid event detection (Hila Becker, Naaman, and Gravano 2010) very few of the tweets that I have collected from Norway contains location information (typically 2-3%). I discarded the time dimension as some topics are continuous, and multiple clusters with the same topic are undesirable for this experiment. As a result of this the clustering must be based on the texts themselves. This differentiates this study from others. While time is very important to make sense of the world, the exclusion of time in the clustering of the tweets allows for topics of low quantity per time unit but with continuous discussion to be gathered.

Clustering Tweets

Clustering as a method is closely related to information retrieval and search. It can also be used for other activities such as browsing (Cutting et al. 1992) and identification of redundant pieces of information (Nezda 2012). My intention was to utilize clustering to expose themes of topics of discussion in a larger dataset.

Hierarchical clustering has the advantage that the number of clusters does not need to be known a priori, and this makes sense in a corpus with more or less unknown content. A disadvantage is scaling, as a matrix containing distance measure between all documents needs to be created, and this is computationally expensive. Other methods such as k-means takes the number of desired output clusters as an input, but do not require a distance matrix. To overcome this problem I used the Buckshot algorithm (Cutting et al. 1992) to initialize k-means with the results from a hierarchically clustered random subset. While speed still is an issue, the memory needed is limited to what can be found in a typical desk- or laptop computer. For the sake of this experiment processing time is not decisive. While in a real world scenario speed is key and a quicker method of clustering is needed.

As distance measure the Euclidian distance was used for the hierarchical clustering and the cosine angle distance was used for the k-means clustering.

The operationalization of the clustering algorithm was done in Python, drawing on the natural language tool kit (NLTK) (Loper and Bird 2002) and the Oslo-Bergen tagger (a grammatical tagger for Norwegian) for word categorization and lemmatization ("The Oslo-Bergen Tagger" 2012). By removing unwanted word categories (assumed less rich in information such as determinatives, conjunctions, pronouns, etc.) revealed by the tagger, vectors were built with the tf-idf value of each word. The tf-idf value is a numerical description of how important a word is to a document in a collection of documents.

Twitter contains a lot of data that's hard to categorize (noise). The messages are short, often with unconventional abbreviations and slang is heavily used. Spelling mistakes are not uncommon. The bag of words model gets weaker as a result of this. Initial results pointed out some further alteration in the algorithm. Tweets shorter than seven words were excluded from the clustering and presented to the user as a single cluster. I also added a short list of stop words. The reduction of the set is done in effort to condense the concentration of significant and meaningful words. Further an effort was put into boosting particularly meaningful words:

Hashtags is a way for authors on Twitter to label a tweet as in a context or topic. This is done by using the number sign (#) as a postfix to any word. Examples can be #Oslo, #politics, #obama, etc. These words I consider as more valuable than others as the give topical information, and I give them a boost by adding a fixed value to the tf-idf previously calculated for this word. The same is done with nouns, verbs and user mentions (identified by the postfix "@" to a single word) with decreasing boost values.

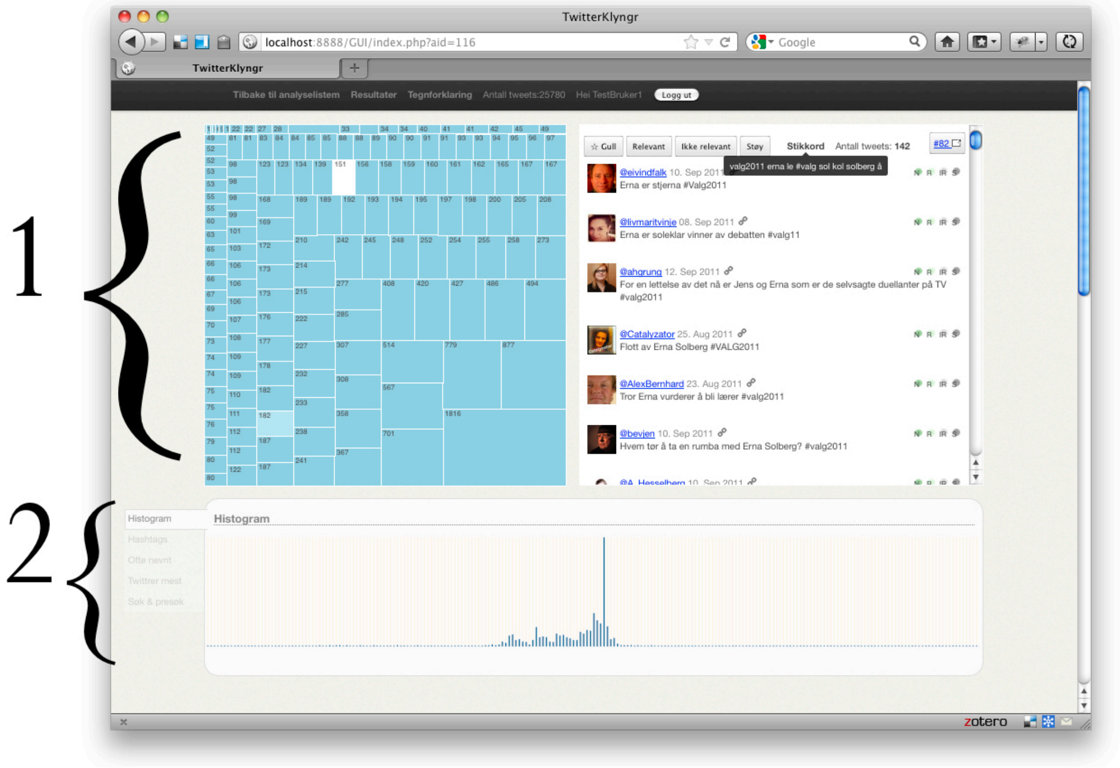
By using the Buckshot algorithm, the number of clusters (k) is decided by the outcome of a smaller initial hierarchical clustering procedure. In the following k-means algorithm a troublesome problem occurs. If a tweet has no or little overlap with one of the existing clusters, it still needs to be put in one. This results in some very large, bloated and inconsistent clusters that offer little aid to understand what is discussed. As a remedy to this I added a breakout mechanism to the k-means; if the distance between a tweet and the nearest centroid of any current cluster is too great, the new tweet is added as a new cluster. The algorithm stops when there are no (or fewer than n) alterations between this and the last iteration. To explore the algorithms ability to cluster material - retweets (that are identical or largely overlapping texts) were removed. Retweets can be fetched back in at a later point, and by doing so the clustering without retweets makes it is clearer to see what texts become clustered together. While retweets are often used as a key factor in understanding Twitter, the exclusion of retweets allows for a clearer view of what different texts that gets grouped together, and also limits the effects of having the same message repeated over and over

A further alteration was done in the presentation of the clusters. To quickly get a sense of what is typical in a cluster I ordered the clusters by the sum of the tf-idf values for each word in a tweet. The clusters were also labeled with the highest-ranking tf-idf word for the individual cluster against all the clusters. The clusters were presented to the participators as rectangles in a one-leveled treemap where the size of the rectangle corresponds to the number of tweets in the cluster (see figure 1). The clusters were labeled with the cluster size, with top key words revealed by hovering over each rectangle. By clicking on a rectangle, the clusters' content is displayed in a side-by viewer.

Flow of the Algorithm

The clustering algorithm starts with input of how many tweets to fetch from the database. The Oslo-Bergen tagger preforms lemmatization and determines word categories, and words from unwanted categories are removed. Text with very few words (n<7) are removed from further processing but kept as a separate cluster (so these tweets sill can be found in searches and though browsing meta data in the application GUI).

Further vectors are created for each tweet, with tf-idf values representing each word per tweet. A boost is added to @mentions, verbs, nouns and hashtags. Through extensive experimentation I ended up using {+ 0.15 for @mentions and verbs, +0.3 for nouns and +0.5 for hashtags}, a set of values that worked well in practice. A distance matrix is then created for a random sample of the vectors and a hierarchical clustering is performed. The centroid of each cluster from the hierarchical clustering is calculated and used as seeds for the k-means clustering. The k-means is performed on all vectors and is ended when zero (or fewer than n if number of iterations is greater than nn). If a tweets has a distance to a current centroid that is greater than 0.009 (where 1 is identical and 0 is no overlap) a new cluster is created from the text.



valg2011 erna le #valg sol kol solberg å

@eivindfalk 10. Sep 2011
Erna er stjerna #Valg2011

@livmaritvinje 08. Sep 2011
Erna er soleklar vinner av debatten #valg11

@ahgrung 12. Sep 2011
For en lettelse av det nå er Jens og Erna som er de selvsagte duellanter på TV #valg2011

@Catalyzator 25. Aug 2011
Flott av Erna Solberg #VALG2011

@AlexBernhard 23. Aug 2011
Tror Erna vurderer å bli lærer #valg2011

@bevjen 10. Sep 2011
Hvem tør å ta en rumba med Erna Solberg? #valg2011

@A_Hesselberg 10. Sep 2011

Figure 1: Screenshot of the graphical user interface. 1 points to a single layer tree graph where each cluster is a rectangle (the cluster display). Hovering each cluster shows the top key words, hashtags and users for this cluster as a tool tip. By on clicking a cluster the tweet in this cluster is displayed in the viewer to the right. On top of the viewer the user can prioritize the cluster (pure gold, interesting, not interesting, noise). Single tweets can be annotated in the same matter (to the right of each tweet). In the end of a session the user can view her annotated clusters and tweets in prioritized order. 2 points to Meta data (histogram, hashtags, users and @mentions by frequency, and search. By clicking on items from the Meta data (a day, a hashtag, a user, etc.) the clusters that contain these tweets are highlighted in the cluster display and the tweets are displayed in the viewer

Data

Data for the study was collected using YourTwapperKeeper (O'Brien III 2010). This is an application for fetching and storing tweets collected from the Twitter API. My application shears database tables with YourTwapperKeeper. The data was collected over 8 months, from May to December 2011 and are search results based on hashtags used in

1) A broad debate: the 2011 Norwegian local election (n=25780, hashtags: #valg, #valg11, #valg2011)
2) A narrow debate: the EU Data Retention Directive (n=6279, hashtag: #DLD).

The narrow debate is a smaller set and was used as a warm-up exercise to let the participants get to know the system before the larger set with election data was examined.

This data was chosen over the potential use of standardized data collections from the text mining community to offer professional Norwegian journalists data that likely is relevant to their day-to-day activities, in their own language.

Study

To gain understanding of how the clustering application and method appeals to journalists an exploratory study was conducted. After a pilot study with students from the local student newspaper a total of seven test-sessions was done with professional journalist. These were selected from national, regional and local media institutions (TV, online and paper). All participants had experience with social media as a source for news production, and all used Twitter in their workplaces. The participants were relatively young (from 26 to 45 years old) and digitally literate, but also relatively experienced in the news business (from 4 to 18 years). The session with each participant lasted about an hour (M = 64 min), including a five-minute introduction and a twenty-minute interview after the participant had used the system. The participants were unpaid.

The test sessions were conducted in the participants local work environment when possible. Five sessions were done in silent rooms connected to the main newsrooms, while two were done in a seminar room at the university campus.

The experimental procedure was an introduction to the system followed by a small dataset for the participants to get to know the interface and functionality. A larger dataset was then presented with the task of 1) evaluating the clustering ability and 2) find material that they thought could be worth reporting or looking further into.

The testers were encouraged to talk allowed during the test session. The audio from both the test sessions and the interviews was recorded and fully transcribed. The researcher also recorded observations.

Criteria for evaluation was the perceived utility of the approach, in qualitative terms, as described by the testers in regard to what they considered to be important when searching for stories, sources or trends in the data.

Findings

Evaluating the clustering ability is an exercise in finding overlap in the algorithmically generated groups, and the participants' expectations to what such groups of sub-theme texts should contain. The search for material worth reporting is a search for "interestingness" and building on the literacy gained from the first task, an exercise in applying journalistic interest to a given dataset. The latter exercise can tell a lot of how the clustering approach should be executed.

Some overarching results include the unfamiliarity with working with Twitter data as an object of analysis; another is the experience that overwhelmingly large datasets are discouraging in spite of being broken down into smaller units (156 and 183 clusters) and; that while the participants had different fields of interest and different ideas of what could be worth looking further into, much of the same focus was put on who the authors of the tweets are and what functions they have in the context they are in. Finding sources was equally interesting to finding good stories or noteworthy arguments.

Deciding where to look – gaining literacy

The bag of words model for natural language processing with a clustering strategy that incorporates a simple distance measure creates clusters that are statistically similar, but occasionally semantically and pragmatically inconsistent. The sentences "I don't hate you, I love you" and "I don't love you, I hate you" are considered equal. While both sentences are related to strong feelings of love and hate, this kind of grouping did lead to some realization of methodological shortcomings.

> *Here is one cluster that has clustered together both school and school election, and election and win. It has clustered some tweets about school elections, and also politics about schools (P6).*

This characteristic becomes a problem when two sides of the same debate utilized the same words to express very different ideas:

> *Here the key words are 'to stop', so there is a lot of 'stop DLD', but also a lot of 'DLD stops'. 'Could DLD have stopped 22/7?' vs 'your donation to stop DLD (P2).*

As much as this is a shortcoming in the chosen methodology it becomes a matter of media literacy in practice. “So it is as much about stopping the Data Retention Directive as if the Data Retention Directive may be to stop anything. So it is both sides of the debate” (P2).

This lead to the question of what kind of texts that would be “hidden” by the noise in clusters with poor key words and unclear topics, particularly in large clusters that the participants considered too big to properly look though:

> *Some of the categorizations - the clusters - are totally uninteresting while some are interesting. So what this actually does is to cluster together some conversations, some discourses that are found on Twitter, so that you do not have to do it yourself. It also categorizes the stuff that isn’t interesting, but you can’t quite know whether to trust it fully, because it only says that all these 1816 tweets doesn’t fit into any other topics. But there might be interesting stuff here; only it hasn’t found any common denominator that they fit (P6).*

The flipside of this feature is that topics that are known can be scattered across multiple clusters and thus diluted. “This story should have had a lot more tweets. The one with the teacher in Kongsberg or wherever, I can’t understand this properly. Why are there only 12 tweets here?” (P5) The expectation of what a cluster should be and what the clusters from my algorithmic clustering are - is slightly different. The participants expected subsets of a debate to be thematically divided or “threads”. Not linguistically similar texts. In spite of this, evaluation of the algorithm shows promise for further tailoring and adaptation of the algorithm and possibly even more important: help identify and highlight clusters that contain texts that are expected to be interesting (e.g. containing known named entities, etc.).

The clusters that did yield most positive feedback were typically small in size and considered clear (absence of noise). These were perceived as more specific and more interesting.

> *It’s the same here, here’s one that has ‘moe’ and ‘borten’ and ‘ola’, ‘bort’ and ‘2011’, so here you have tweets about Ola Borten Moe, that suggests that by skimming though this cluster –with 63 tweets – we can say something about what people think of him in this data. Generally, by looking at smaller clusters, you get more specific key words (P6).*

Although the datasets were big and the interest fields of the participants are different, some clusters were identified and commented by multiple testers.

> *Here it has come across something that has to do with first times. So if you look for first-time voters, this cluster would be very interesting (P2).*

This cluster did not gravitate towards a named entity and functions as an example of how this methodology can construct clusters concerning concepts simply through the similarity of wording.

> *This is something I could have checked out to see if I wanted to write a story about; someone that has voted for the first time. ‘Looking forward to vote for the first time tomorrow’. Perfect, we give him a call to ask if we can join him. It is like this: you can show up at the polling station, but then you risk going to the wrong station, or maybe you can’t find any first-time voters when you’re there, or they say no (P4).*

The overall experience with evaluating clusters to identify where interesting chunks of a debate are to de found in this tool, all results point towards purity. The lack of noise and immediate consistence (strong signal) is good, and this was found in smaller clusters that tend to have more specific key words.

Divide and Conquer

The prototype offers simple methods to extract and highlight content based on various meta-data (dates, hashtags, mentions) and free text search. The ability to save clusters or tweets to a user-specific list also strengthens the focus on encompassing and including subsets the user finds interesting or pertinent. The participants had no complaints concerning this, but through use the opposite function – exclusion – were found lacking. Just what to exclude varied with the participants (dis)interest.

Some found particular authors noisy: “To opt out of threads that @nicecap participates in would have help a lot" (P1). Who to pay attention to and who to exclude is important.

> *If you can remove content, create a Twitter group of media people and politicians, and then search for immigration... […] don’t get me wrong, the ‘important’ people must be abstracted (P7).*

While others would like to remove geographically bound clusters: "Rana, election in Rana, I'm not interested in that. Trondheim, I'm not interested in that" (P3). To be able to remove subsets based on meta-data was also demanded.

> *This is clearly something Swedish I am not inn on. #Acta here seems like something Swedish. Høyre [the conservative party] is Norwegian .. Yes. #FRP [the progress party], #AP [labour party], #EU [the european union]. #2pl - this is some football stuff? Is it possible to remove tags from the dataset? Or mark this #2pl as irrelevant for instance? (P3).*

During testing it became clear that while the testers did have special fields of interest and ideas on how to find data in their field, the massive amounts of data invites to explore and investigate outside of their daily topical spheres. When clusters and subsets (e.g. by hash tag) were identified as related to something they recognized (e.g. media stories; events; persons; or locations) they had the need to exclude it to see what is left when this is removed. The focus on finding the interesting is also a matter of removing what is known and uninteresting, and this should be included in the requirements list for a future tool.

Finding Stories

The initial idea for the tool was to aid journalists in finding stories; follow-up stories and sources in social media, or to confirm that the journalist has a fair overview of a debate. When asked to identify trends, tweets or other noteworthy findings that could be used for stories in their workplace, the participants all found something to show. This does not mean that these findings would actually end up in print; broadcast or web media, the unaccustomed setting of a user test session, with a researcher present, does skew these story-findings, but the results might illuminate what sort of stories this kind of tool does inspire.

The already mentioned first-time voter angle is a story that is regularly told in relation to elections, other findings the participants found are similar to this in regard that they are stories that often is covered: property taxes (P4), voters reactions after voting and election vigils (P3). Another category are Twitter-related stories such as small political parties that has few votes but that generate much interest and debate on Twitter. Among other Twitter-related angles was the distribution between the national party leaders and parties in mentions and activity, and while this was commented upon frequently, the seemingly predictable results (the distribution looks a lot like the election result) was not considered worth reporting upon.

> *It’s just the usual suspects here. It’s definitely the political left that is most active in social media. That is almost a story. Which [parties that] are mentioned most. It is red-green. Perhaps it’s an effect of the current government; it’s hard to tell. But it is interesting never the less (p3).*

The stories that were identified while browsing are mostly curiosities that are immediately surprising or extraordinary. "That looks like a story, if there is a 94 year old man that is going to the municipal council, that is a story" (P3). The fact that journalists are looking for stories in the margins of normality were underlined repeatedly.

> *This is an odd cluster –with the key word "go"/"come on". It indicates… it is almost like someone sits in a sporting arena and cheers at the debate. It spans from 'go, go, go' and nothing more to someone cheering for the TV hosts to someone cheering for a party. This could be a curiosities-story (P6).*

A small entertaining surprise does not necessary suffice on an election night though:

> *This is a story. It's witty, if Ferjelista [the ferry list] .. it's dead funny. 'Ferjelista is likely to have two representatives in Volda municipial council'. On an election night there are a lot angles, but on general grounds that is one witty piece of information, but it depends if we would have time to make it (P3).*

The same participant identified a situation where a politicians' (Oddny Miljeteigs') gesticulation on TV was freely interpreted by the Twitter audience as sign language and entertainingly spread as simultaneous interpretation.

As an exception to light-hearted and soft-news-type findings one untold politics story of regime critic was identified: "Sad that deceased from the 22/7 are not removed from the ballots for AP (labor party) in Oslo. This is actually a big story" (P1).

User generated content (UGC) is found to add to the soft- and human-interest news in other studies where user can contribute directly to the media organizations (Deuze, Bruns, and Neuberger 2007; Harrison 2009). My findings indicate that journalists find the same kind of news categories also when the data isn't handed in to them, but when they are looking for stories in UGC.

Finding Sources

The facts that debates on Twitter allows for anyone to share their ideas and comments is received with a certain ambivalence by the informants. On the one hand the democratic aspects and openness for new potential voices are welcomed in warm terms. On the other hand, the need to identify "the usual suspects" is complicated by the share number of voices. Who the authors of tweets are, is of paramount importance.

> *You gain insight into what people care for, you do. And what parties they talk about. It is interesting that so many talk about Venstre [liberal party]. But if you are looking for sources - if you are to get hold of sources - I would not have picked some random person (P7).*

Consequently familiar logos, organizations and persons were appreciated. "It is very pleasant when they show what the stand for [points to an avatar with a political logo]" (P7). While authorities were identified with quotable insights, some positions can be excluding;

> *Thor Bjarne Bore is an editor in a newspaper, so there is no point using him as a source (P1).*
>
> *There is a lot of official stuff here, that is boring. [Interviewer asks: Official?] Yes, official accounts [points to a profile of an NGO] (P3).*

Profiles of politicians and political parties were considered quotable, while the average Twitter user were approached with a very different caution, and whether they could be quoted was taken into a much more thorough consideration.

> *It is also a discussion of press ethics, to what extent we can… -it is something we discuss continuously – whether we can fetch material from social media and incorporate it into stories. I think that what people say on Twitter is public, I do (P5).*

The possibility to waste time and efforts on someone that didn't deliver a strong enough case was clearly expressed.

> *It is relevant to find who this is; I wouldn't use just any person sitting at home in his bedroom being angry at something. But this one there is a systems developer and*

> *information flow geek, liberalist, Dag B, a blogger. That means that he works with relevant things in this matter (P7).*

One participant had come up with a strategy to cope with this issue, and kept track of no-authoritative "potentials" by following them for a while before potentially making contact for a story. "It has happened a few times, so it shows that it is possible to find people that are sources, and not authoritative sources" (P6).

As sources, the non-professional commentator needs to have more than an interesting comment to fit the journalists' criteria: "the most important thing to separate is those who speak about something that happens in the media, and those who speak about something truly unique" (P1). Identifying who authors of Twitter texts are is important to a journalist for many reasons, independent of how clever or interesting the content of the tweet is. A key reason for evaluating tweets so sternly in context of their authors is the journalistic practice of accountably, related to public sources in checking for spin and with civilian sources in checking for competency, relations, credibility, etc.

Conclusion and Further Work

The approach presented in this paper has some limitations with regards to noise and ways to reveal a clusters' content quickly, but results also indicate that cluster analysis can aid the analysis of Twitter messages. The usefulness is related to the quality of the clusters, and the quality of the clusters is a matter of meeting the journalists' expectations.

Contribution: Improving clustering for tweets

Clusters of texts are considered good if they are immediately recognized as related to something the journalist already knows and there is a strong coherence within the cluster. Known entities should have a high priority both on order to examine these in particular, but also in order to exclude them. Ways to improve the clarity (signal) in the clusters are needed.

One idea for a better result is to crowd-source or manually construct the initial clusters before running k-means; another is to base the k-means only on data that contains named entities or other textual favorable features. Who people are is very important to the journalists in this study, and it is reasonable to assume this as an occupational trait. As such, the idea of finding an exceptionally clever insight or argument from anyone regardless of their position in society must come second after the journalistic need to anchor voices to positions and groups of people. This can be potentially be operationalized by basing the initial centroids on texts that contain such entities.

Other more crude approaches that exclude material could also be applied (to exclude texts based on the lack on machine recognizable entities, frequency of spelling mistakes, text length, etc.). This might increase the experienced utility, but adds a bias towards particular groups of authors and dims the democratic aspects of using social media as sources.

Elements that worked well in this study is the general idea of reducing the amount of units a journalist needs to examine to get an overview of twitter data, browsing by metadata and using tf-idf to label and order sub-sets of data. The deliberate omission of the time aspect in the analysis worked without any problems in this study. The histogram that was offered for exploration obtained the chronological order and amount of texts in time. Improvements in the application include the ability to exclude data (from persons, hashtags, etc.), navigate by user types (politicians, media-people, organizations, etc.) and producing an even clearer signal.

In spite of the democratic promise of social media where anyone can participate, who the authors of messages are is very important to journalists. The “usual suspects” for journalists are people of power (politicians, organizations, celebrities and experts of various kinds), and in the category for “John citizen” they are looking for “cases”; persons that can exemplify a bigger phenomenon. Who the authors of tweets are were examined when tweets of interest were found and is too important for journalists too not include in the design of an analytics tool. Prior research projects have identified roles of Twitter authors though directories (An et al. 2011) and such lists could be used for filtering in an analytics tool for journalists. A similar approach was used to identify tweets that link to established media institutions (N. Diakopoulos, De Choudhury, and Naaman 2012) an approach that can be implemented to support the need to filter out opinions that is likely to be third hand information.

Acknowledgments

I will direct a special thanks to the NRK- team for valuable feedback throughout this project, and Matthew Weinstein for modifying TAMS Analyzer to support Scandinavian languages – an valuable improvement for doing the analysis of the data in this study.

References

An, J., M. Cha, K. Gummadi, and J. Crowcroft. 2011. “Media Landscape in Twitter: A World of New Conventions and Political Diversity.” *Proc. ICWSM* 11.

Becker, H., M. Naaman, and L. Gravano. 2011. “Selecting Quality Twitter Content for Events.” In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM’11)*.

Becker, Hila, Mor Naaman, and Luis Gravano. 2010. “Learning Similarity Metrics for Event Identification in Social Media.” In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 291–300. WSDM ’10. New York, NY, USA: ACM. doi:10.1145/1718487.1718524. http://doi.acm.org/10.1145/1718487.1718524.

De Choudhury, M., N. Diakopoulos, and M. Naaman. 2012. “Unfolding the Event Landscape on Twitter: Classification and Exploration of User Categories.” In *Proc. CSCW*.

Cutting, Douglass R., David R. Karger, Jan O. Pedersen, and John W. Tukey. 1992. “Scatter/Gather: a Cluster-based Approach to Browsing Large Document Collections.” In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 318–329. SIGIR ’92. New York, NY, USA: ACM. doi:10.1145/133160.133214. http://doi.acm.org/10.1145/133160.133214.

Deuze, Mark, Axel Bruns, and Christoph Neuberger. 2007. “PREPARING FOR AN AGE OF PARTICIPATORY NEWS.” *Journalism Practice* 1 (3): 322–338. doi:10.1080/17512780701504864.

Diakopoulos, N., M. De Choudhury, and M. Naaman. 2012. “Finding and Assessing Social Media Information Sources in the Context of Journalism.” In *Proc. Conference on Human Factors in Computing Systems (CHI)*.

Diakopoulos, N., M. Naaman, T. Yazdani, and F. Kivran-Swaine. 2011. “Social Media Visual Analytics for Events.” *Social Media Modeling and Computing*: 189–209.

Diakopoulos, Nicholas, Munmun De Choudhury, and Mor Naaman. 2012. “Finding and Assessing Social Media Information Sources in the Context of Journalism.” In *Proc. Conference on Human Factors in Computing Systems (CHI)*. http://research.microsoft.com/en-us/um/people/munmund/pubs/chi_2012.pdf.

Futsæter, Knut-Arne. 2012. “MedieTrender 2011” February 12. www.tns-gallup.no/arch/_img/9100748.pdf.

Harrison, Jackie. 2009. "USER-GENERATED CONTENT AND GATEKEEPING AT THE BBC HUB." *Journalism Studies* 11 (2): 243–256. doi:10.1080/14616700903290593.

Hermida, Alfred. 2010. "TWITTERING THE NEWS." *Journalism Practice* 4 (3): 297–308. doi:10.1080/17512781003640703.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. "Why We Twitter: Understanding Microblogging Usage and Communities." In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, 56–65. WebKDD/SNA-KDD '07. New York, NY, USA: ACM. doi:10.1145/1348549.1348556. http://doi.acm.org/10.1145/1348549.1348556.

Kwak, Haewoon, 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*, 591–600. WWW '10. New York, NY, USA: ACM. doi:10.1145/1772690.1772751. http://doi.acm.org/10.1145/1772690.1772751.

Larsson, Anders Olof, and Hallvard Moe. 2011. "Studying Political Microblogging: Twitter Users in the 2010 Swedish Election Campaign." *New Media & Society* (November 21). doi:10.1177/1461444811422894. http://nms.sagepub.com/content/early/2011/11/21/1461444811422894.

De Longueville, Bertrand, Robin S. Smith, and Gianluca Luraschi. 2009. "'OMG, from Here, I Can See the Flames!': a Use Case of Mining Location Based Social Networks to Acquire Spatio-temporal Data on Forest Fires." In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, 73–80. LBSN '09. New York, NY, USA: ACM. doi:10.1145/1629890.1629907. http://doi.acm.org/10.1145/1629890.1629907.

Loper, Edward, and Steven Bird. 2002. "NLTK: The Natural Language Toolkit." *arXiv:cs/0205028* (May 17). http://arxiv.org/abs/cs/0205028.

Luo, Zhunchen, Miles Osborne, and Ting Wang. 2012. "Opinion Retrieval in Twitter." In *Sixth International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewPDFInterstitial/4592/5044.

Nezda, Luke. 2012. "ClusteringInDepth. Methods and Theory Behind the Clustering Functionality in Google Refine." Accessed March 6. http://code.google.com/p/google-refine/wiki/ClusteringInDepth.

O'Brien III, John. 2010. *Your TwapperKeeper – Archive Your Own Tweets - Http://Your.twapperkeeper.com*. http:// your.twapperkeeper.com.

Shamma, D. A, L. Kennedy, and E. F Churchill. 2010. "Conversational Shadows: Describing Live Media Events Using Short Messages." *Proceedings of ICWSM*.

Shamma, David A., Lyndon Kennedy, and Elizabeth F. Churchill. 2009. "Tweet the Debates: Understanding Community Annotation of Uncollected Sources." In *Proceedings of the First SIGMM Workshop on Social Media*, 3–10. WSM '09. New York, NY, USA: ACM. doi:10.1145/1631144.1631148. http://doi.acm.org/10.1145/1631144.1631148.

Stray, Jonathan. 2012. "The Overview Project." Accessed April 30. http://overview.ap.org/.

"The Oslo-Bergen Tagger." 2012. Accessed March 7. http://www.tekstlab.uio.no/obt-ny/english/index.html.

Tumasjan, A., T. O Sprenger, P. G Sandner, and I. M Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment." In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 178–185.

Weng, Jianshu, and Bu-Sung Lee. 2011. "Event Detection in Twitter." *Proc. of ICWSM*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2767/3299.