Applications of Opinion Mining to Data Journalism

Jacopo Ottaviani

Dissertation for obtaining the Master Degree in

International Masters in Human Language Technology and Natural Language Processing

15th of April 2013

Advisor	Professora Paula Cristina Quaresma da Fonseca Carvalho (INESC-ID Lisboa and & ISLA Campus Lisboa | Laureate International Universities)
Co-Advisor	Professor Michael Thelwall (University of Wolverhampton)

Index

Resumo

Os media sociais desempenham um papel fulcral na sociedade da informação. Um grande volume de dados gerados pelos utilizadores gira em torno das redes sociais disponíveis, como o *Twitter*, tendo um importante impacto na indústria de comunicação e no dia-a-dia dos próprios indivíduos. O uso crescente e massificado das redes sociais por parte dos utilizadores, facilmente acessíveis através de qualquer computador pessoal ou telefone inteligente, resulta da necessidade constante de partilha de informação e expressão de opinião acerca dos factos que os rodeiam.

Nos últimos anos, as áreas de processamento de linguagem natural e de análise de sentimento têm procurado desenvolver técnicas e tecnologias capazes de analisar e extrair dados sobre notícias que circulam nas redes sociais. Ora, as aplicações de prospeção de opinião orientadas para análise de conteúdos gerados pelos utilizadores, em particular *tweets* com ligação a notícias, podem fornecer novos pontos de vista sobre o comportamento do público geral a factos noticiosos e ajudar a interpretar a evolução dos sentimentos, face a esses factos. Em particular, a análise das referências a notícias nas redes sociais permite (i) medir o impacto que as notícias têm sobre os leitores e (ii) agregar elementos que contenham histórias em comum.

Numa perspetiva mais ampla, o objetivo principal desta tese consiste em demonstrar como a prospeção de opinião (ou análise de sentimento) pode ser adotada no âmbito do jornalismo computacional. Com este trabalho, esperamos poder contribuir para a criação futura de uma ferramenta de análise de sentimento capaz de responder a perguntas como, por exemplo, "Quais são as notícias que desencadeiam as reações mais positivas ou mais negativas junto dos leitores?" ou "Quais são as notícias que provocam os maiores contrastes de opinião na comunidade das redes

sociais?", e de visualizar as emoções dos utilizadores face a notícias relacionadas com entidades ou eventos específicos, numa perspetiva espácio-temporal.

Concretamente, a nossa pesquisa tem por objetivo investigar a forma como o *SentiStrength* - uma ferramenta de análise de sentimento especificamente concebida para o processamento de conteúdos gerados pelos utilizadores, em particular, mensagem informais breves - pode ser adotado e otimizado na deteção de sentimento em *tweets* associados a notícias. Em particular, este trabalho pretende responder a três perguntas de investigação, que enunciamos em seguida. Primeiro, como é que o *SentiStrength* se comporta face a *tweets* que contenham ligações para notícias? Segundo, será que o *SentiStrength* classifica melhor as mensagens que incluem o título da notícia e respetivo comentário ou, por outro lado, será que o processo de classificação é mais preciso se a análise se cingir apenas ao comentário expresso nos *tweets*? Terceiro, qual o impacto das diferentes polaridades de opinião nos casos anteriormente descritos?

O corpus que construímos no âmbito da nossa pesquisa é constituído por aquilo a que chamamos de "tweets com ligação a notícias", isto é, trata-se de tweets que incluem um URL de um artigo, o título notícia a que se refere esse URL e, finalmente, um comentário sobre essa notícia. Depois de coligidos os dados que obedeciam a esta estrutura, a partir do site *The Guardian*, estes foram pré-processados. Posteriormente, foi criada uma coleção dourada, utilizada para a avaliação do desempenho de *SentiStrength*, em particular de forma a dar resposta às perguntas de investigação previamente formuladas.

Os "tweets com ligação a notícias" são interessantes por vários motivos. Em primeiro lugar, é interessante perceber de que forma é que o desempenho do *SentiStrength* pode ser afetado pela presença de *tweets* que incluam quer informação factual (o título da notícia) quer informação de cariz subjetivo (o comentário à própria notícia). Por outras palavras, é importante perceber se o título pode ser uma fonte de informação relevante ou uma fonte de entropia na análise de opinião.

Segundo, é igualmente útil perceber se os utilizadores são mais diretos na expressão de opinião, quando confrontados com um limite de carateres mais reduzido, melhorando, assim, o desempenho global do *SentiStrength*. Terceiro, é importante perceber se a análise global do título da notícia e respetivo comentário poderá estar dependente da polaridade da opinião envolvida.

Os resultados das experiências que levámos a cabo neste trabalho permitiram-nos concluir que o *SentiStrength* tem um melhor desempenho em termos de classificação de sentimento, se omitirmos os títulos dos "tweets com ligação a notícias". Em termos globais, a medida-F aumenta 0.32 quando a informação do título é descartada no processo de análise de sentimento.

Os resultados obtidos mostram, ainda, que a deteção de polaridade positiva e neutra dos "tweets com ligação a notícias" é amplamente melhorada quando o título é excluído da análise. Porém, no caso das mensagens negativas, observámos que a polaridade é detetada de forma mais precisa se os *tweets* incluírem, além do comentário, o próprio título.

Neste contexto, propomos um pseudo-código de um procedimento para incluir no *SentiStrength*, de forma a lidar mais eficazmente com o tipo de dados que exploramos nesta tese.

Abstract

Nowadays social media play a central role in every day life. A huge volume of user-generated data spins around online social networks, such as *Twitter*, having an extraordinary impact on the media industry and on the users' everyday life. More and more users and people use social networks from their computers and smartphones to share their emotions and opinions about the facts happening in the world. Natural language processing and, in particular, sentiment analysis are key technologies to make sense out of the data about news that circulates in the online social networks. The application of opinion mining to news-oriented user-generated contents, such as news-linking *tweets*, can provide novel views on the news audience behaviour and help to interpret the evolution of sentiments. Applying this capability in the social news-sphere permits (i) to measure the impact of news onto readers and (ii) to gather elements that contain stories.

From a broad perspective, the main aim of this research is to face this challenge, that is, to explore how opinion mining (or sentiment analysis) can be adopted into the field of digital media and data-driven journalism. Having an accurate sentiment analysis tool working on a large-scale corpus of news-linking *tweets* would allow answering ambitious questions like "What are the news that trigger the most positive or negative reactions?" or "What are the news that provoke the biggest contrasts among the social network community?" or to see how the global emotions about news-related entities change on time and how they are distributed on the space (by exploiting the geolocation of *tweets*) — we are not answering to these questions, but we move towards this direction. In particular, we do this by testing and improving a sentiment analysis tool in the news comments domain.

Concretely, our research aims to investigate how *SentiStrength* – a sentiment analysis tool for short informal text – can be adopted and optimized to detect sentiments out of news-linking *tweets*. We have depicted and answered to three research questions. First, how does *SentiStrength* perform with news-linking *tweets*? Second, does *SentiStrength* classify more accurately the polarity of news-carrying *tweets* if it considers the news-title and the comment, or if it considers only the comment? Third, how do the different polarities singularly perform in the above-mentioned cases?

The specific type of *tweets* we gather in our corpus is what we call "news-linking *tweet*", that is, a tweet that includes a URL to a news article, the news title and a free textual comment about it. We test the performances of *SentiStrength* when dealing with this particular data structure, in particular with *tweets* that include a link to an article from *The Guardian* website. In order to address this, we set up an experimental process that started from the data collection and pre-processing, continued with the creation of a gold standard and ended with interpretations of *SentiStrength*'s outputs and led to the answers of our research questions.

The news-linking *tweets* are interesting for several reasons. First, it is worthwhile to consider to what extent the titles decrease *SentiStrength*'s performance when dealing with news comments. In other words, when dealing with news-linking *tweets*, is the title a source of entropy or a relevant source of information? Second, it is interesting to see if the reduction of space available for the comment reduced also the presence of ironic or ambiguous comments, improving consequently its overall performance. Third, by having both the title and the comment, it is interesting to combine those elements in the different cases of positive and negative polarities. This way it is possible to see whether or not the title in some cases is essential for a better sentiment classification.

The whole experimental process led us to the conclusion that *SentiStrength* better detects sentiments in news-linking *tweets* that do not include titles, but only include user comments (what we call *clean tweets*). *SentiStrength* reaches, in fact, a better performance both in terms of Precision, Recall and f-Measure when it receives clean *tweets* as input. In particular, the average f-Measure is 0.53 (+0.32 higher of the f-Measure obtained with the *complete tweets*).

Our results also include further findings, showing how in the specific case of sentimentally positive or neutral news-linking *tweets SentiStrength* performs better if it excludes from the input the news titles (i.e., only comment). This does not hold with negative news-linking *tweets*, which are better detected if inputted to SentiStrength in their complete form (i.e., title and comment). Our research provides, finally, the pseudo-code of a procedure for *SentiStrength* to deal with news-linking *tweets*.

Besides the experimental results, the corpus we built to answer our research questions allowed us to lay the foundation of a wider platform for sentiment analysis applied to the news sphere. By analysing the data we have gather it will be possible to focus on readers' reactions to news and visualize it in various forms and inspiring new data journalistic applications. In fact, such platform automatizes a series of steps to play with *Twitter* data, from data collection to sentiment analysis, and opens the doors to applications that take advantage of the relationships and connections between *Twitter*, online-news and opinion mining.

Palavras-Chave
Keywords

Palavras-Chave

Prospeção de Opinião, Analise de Sentimento, Processamento de Linguagem Natural, Jornalismo Computacional, Jornalismo de Dados

Keywords

Opinion Mining, Sentiment Analysis, Natural Language Processing, Computational Journalism, Data Journalism

Chapter 1
Introduction

The main aim of the thesis is to explore the connections between the online news-sphere (i.e., the digital media) and the social networks. In particular, we investigated the application of opinion mining (or sentiment analysis) to the news sharing and commenting process across the social networks. Namely, we needed a social network, a news outlet and a sentiment analysis technology. We chose the following: as a news outlet we selected the website of *The Guardian*[1], a British daily newspaper; as a social network we chose Twitter[2], a growing news-oriented social network where users share short text messages consisting of a maximum of 140 characters; finally, as an opinion mining technology we chose SentiStrength (Thelwall *et al.*, 2010), a sentiment analysis tool for short informal texts.

SentiStrength was already tested with generic Twitter corpora (Thelwall *et al.*, 2010 and 2012). However, it has never been tested considering the distinction between *tweets* involving the news domain, i.e., with news-links in them. Since we want to explore how sentiment analysis can be used in the online journalism sphere, and in particular, to analyse how readers comment and react to news online, we take in account this distinction: we consider *tweets* that include news links, news titles and free text comments (presumably) referring to them. We define "news-linking tweets" this particular class of *tweets* that includes a link (i.e., a URL) to an online news article, the title of the news itself and a text comment to that.

[1] The Guardian is a British daily newspaper. Its website is http://www.guardian.co.uk and contains nearly all of the content of the newspapers The Guardian and The Observer, as well as a body of web-only work produced by its own
[2] A presentation of Twitter can be found in the article "What is Twitter?" released by the company and available at https://business.twitter.com/en/basics/what-is-twitter/

A generic news-linking tweet has the following form:

<news title> <news link> <user comment> <fixed mention>

The order of the fields is not necessarily this, although it often comes in this form. An example of a news-linking tweet follows:

Sewing cafe opens in Paris http://gu.com/p/2gc53/tw via @guardian. I want to open one of these in America.

In the example, the news title is "Sewing café opens in Paris". The news link is the URL that points to the article by *The Guardian*. The fixed mention is "via @guardian". And the comment, in this case expressing a positive sentiment polarity, is "I want to open one of these in America."

The length of such data structure – as all the other tweets – is maximum 140. Thus, the number of characters available for the *<user comment>* field is lower than the generic tweets' length. This is one of the reasons that make interesting the investigation of news-linking tweets. It is interesting to see to what extent a reduced space for comments makes SentiStrength work differently than the usual. This, in other words, allowed us to put forward the hypothesis that disposing of less characters, there would be less room for irony. Irony, in fact, is shown to be one of the major issues that affect the performance of SentiStrength (Thelwall *et al*., 2012).

We make the assumption that the comment included in a tweet is referring to the news that is linked, and we only deal with English news and comments.

The research questions of the master thesis can be summarized as follows:

1. How does SentiStrength perform with news-carrying tweets?

2. Does SentiStrength classify more accurately the polarity of news-carrying tweets if it considers the news-title and the comment, or if it considers only the comment?

3. How do the different polarities singularly perform in the above-mentioned cases?

We have depicted three research questions. However, the objective of the master thesis is two-fold and goes further. First, as described in the research questions, it is aimed to test the technology of SentiStrength with a particular class of tweets, namely, news-linking tweets. Second, and more broadly, it is aimed to explore the possible applications of sentiment analysis into data driven journalism. By producing a corpus with news-linking tweets we will provide a concrete framework to let new stories emerge from Twitter data. In order to address this two-fold objective, we provide an introductory review of the key areas up to their state-of-the-art.

The key areas that we are describing concern sentiment analysis, from both a general and a news-focused point of view – presenting the most important technologies currently adopted to achieve the news-oriented sentiment analysis task. In the second part of the literature review we will focus on SentiStrength, which is the technology we are going to test with the news-linking tweets. Next, we provide some definitions related to opinion mining and sentiment analysis that will be used to annotate our dataset and evaluate the outputs of SentiStrength. The fourth part shows articles that describe the central role of Twitter in the news-sphere, depicting thus the rationale behind the choice of our data source. The last part of the literature review

consists of a brief introduction to data driven journalism, describing the genesis of this approach to journalism and its evolution.

Chapter 2
Literature Review

2.1 Sentiment Analysis and News-related Sentiment Analysis

As regards opinion mining in general, Liu (2012) released a book covering many important topics and subtopics related to sentiment analysis and opinion mining. The book provides all the basic concepts necessary to understand the theory and the applications of sentiment analysis in various contexts, spanning from document sentiment classification to subjectivity detection, from opinion summarization to opinion spam detection.

Another detailed survey available in literature is the monograph by Lee and Pang (2008). One major problem of the survey is its anachronism: recently sentiment analysis, as well as the whole Internet Technology, moved fast-forward. Social networks, which are nowadays widespread, do not appear among the mentioned domains of application. Sentiment analysis of short informal text is not covered. However, most concepts, theories and techniques are still valid. The authors also tackle the subtask of opinion extraction between discourse participants. From (Agrawal, 2003) emerges that users who respond to comments in newsgroups tend to be antagonistic with each other. Precisely, about 74% of responses were found to go against what previously said and 7% only reinforce it. Moullen and Malouf (2008) analyse “quotes” in comments. Retrieving comments from politics.com they show that most quotes come from opposed political factions who quote each other instead of their representatives. The research is domain-dependant and its conclusions cannot be extended to Twitter: as Boyd et al. (2010) show, retweets are diverse in purpose. A

list of ten reasons for why people retweet is showed. These reasons span from endorsement to validation, from personal homage to bookmarking, from self-gain in terms of visibility to simply spread tweets to followers.

Thelwall et al. (2010, 2012) proposed SentiStrength, a tool for sentimental analysis designed for short informal English text. The authors design their tool in order to detect positive and negative polarities of short, informal, ideally domain-independent, English text. SentiStrength includes both supervised and unsupervised approaches (see Chapter 2.3).

Balahur *et al.* (2010) realised that in order to define a new methodology for "sentiment analysis of news articles", the task needs to be split into various subtasks. First, target detection that consists of spotting out on what 'object' the article is focusing on. News' targets are usually wider than reviews' targets (e.g., products) and span larger domains; second, the necessity of distinguishing good/bad reported *news content* from good/bad-expressed *sentiments*. In fact, previous experiments showed how machine learning algorithms associated "bad sentiments" to words included in "bad news" contexts, such as the 2008 financial crisis or the Palestinian-Israeli conflict, even if the authors were not conveying any sentiment by themselves; third, sentiment analysis of "world-knowledge-free" opinions, i.e., those opinions that do not involve any direct connection with facts of the world. Furthermore, while addressing news sentiment analysis, three points of view can be spotted: author, reader and text – each of which is supposed to be addressed separately. Regarding this last observation, their conclusion is that sentiment analysis should focus on the *text* level – as *authors* and *readers* have personal backgrounds that influence their views and cannot be predicted.

News-related opinion mining is object of research by the INESC-ID laboratory. A platform named REACTION[3] (Retrieval, Extraction, and Aggregation Computing Technology for Integrating and Organizing News) is currently under development, aimed to gather computational methods for journalism. Despite it is mostly based on

[3] The REACTION project is an initiative for developing a computational journalism platform (mostly) for Portuguese. It can be found at http://dmir.inesc-id.pt/project/Reaction

Portuguese, some works are not strictly language-dependant. Carvalho et al. (2010), for instance, tackle news-related opinion mining in political news. In order to build a corpus to train machine learning sentiment analysis supervised algorithms, the authors (i) collect manually an amount of opinionated user-generated comments on politics, (ii) define a set of syntactic-lexical rules (supported by a sentiment lexicon) in order to detect sentiment polarity (negative/positive opinions) of the comments, (iii) propagate detected polarities to other sentences in the corpus that mention the same "entities". Precision rates reached ≈90% and 60% for negative and positive polarities, respectively. Propagation worked perfectly for negative comments (≈100% success) and quite well for positive ones (70% success). Performances dropped down for positive comments mostly due to irony and lack of world knowledge. Although they work on Portuguese data, their strategy can be applied to other languages, English included – and the rule set can be adapted as well.

Another methodology related to large-scale sentiment analysis for news can be found in Godbole *et al* (2009). The authors set up their news sentiment analysis framework upon *Lydia* – a news analysis framework published by Lloyd *et al* (2005). Lydia is aimed to build a relational model of named entities (namely people, places and companies) through NLP, co-locations and frequency analysis. Lydia pipeline includes five major phases: (i) a crawling step to retrieve the parsed article texts; (ii) a Named Entity Recognition phase; (iii) juxtaposition analysis during which each entity's co-occurrences are calculated; (iv) synonyms identification, in order to compact those named entities represented by different n-grams; (v) temporal and spatial analysis. The most interesting phase, from a data-driven journalistic approach, is the fifth one. This in fact provides a two-dimensional view of public opinion formation dynamics, analysing newspapers' influence spheres and their evolution. Anyway, being from 2005, when there were no widespread social networks, Lydia does not take in account any social data. This makes the framework out-of-date. For example, through Twitter geolocated shares or using the social network analysis

provided by services as bit.ly[4], as can be found for example in the The Guardian Datablog's map of online news spatial diffusion in Britain (Guardian News and Media, 2012), a better spatial mapping could be performed in order to overcome this.

Godbole *et al* (2009) added a sentiment analysis layer to Lydia. This layer can be logically broken down into three sub layers: (i) Sentiment dictionaries construction (ii) Sentiment index formulation (iii) Significance evaluation. The first step is based on a list of notorious positive and negative adjective (seeds), expanded using Wordnet antonyms and synonyms paths. The resulting dictionary generated contained 18,000 words. Comparing their automatically generated lexicon to a human-annotated one, they perform an average recall of ≈71% and precision ≈90%. The second step is based upon co-occurrences analysis, that is, polarised words are associated to named entities when they appear within the same sentences (separated by period characters); the authors do not use sophisticated parsing techniques as they are concerned by the system's speed. They opt for running a responsive web service at the expense of its algorithms' accuracy. Thus, complex sentences might not correctly score. On the other hand, negations are considered, duplicate articles are ignored and some co-reference resolution (in particular, pronoun resolution) is performed, but authors do not provide details about what algorithms are adopted. Other types of co-reference are partially handled: for instance, *George Best* and *G. Best* are correctly associated under the same entity, but *George Best* and *Belfast Boy* are not. Two indexes are introduced: *polarity* and *subjectivity*. The first indicates percentage of positive opinion associated to an entity; the second measures the quantity of sentiment (of any polarity) raised by an entity – that is a sort of *fuss* measure. Although general evaluations about how these indexes work are not provided, tables with examples are shown. From those emerges a meaningful view on negative and positive entities, at least according to the Western culture.

Godbole et al (2009) conclude their research showing a keen interest in news mapping, both recalling (Mehler *et al*, 2006), another paper related to Lydia, and

[4] Bit.ly is a URL shortening and bookmarking service. It is used on social networks to share URLs in a shortened version. It is available at http://www.bit.ly

hypothesizing further visualisations, in particular sentiment-related maps. This applied attitude makes their work particularly suitable for data journalism.

Mehler *et al* (2006) proposes a strategy to *heat map* news contents over the United States geographic area. The authors scraped 800 daily newspapers in the United States and approximately 300 English speaking newspapers overseas, on a daily basis for a period of time, and mapped the named entities appearing in them onto the US, in form of heat maps. The more heat, the more an entity is associated to an area.

Going through their pipeline, after the crawling phase there is an interesting module of duplicate and near-duplicate articles filtering. In order to detect duplicates, the authors run the plagiarism detection algorithm by Schleimer *et al* (2003) onto their dataset. This way, the authors excluded 190,000+ duplicates and near-duplicates, out of 253,523 downloaded articles. Next step in the pipeline is what authors call "sphere of influence" modelling: by combining (a) newspapers' readership (naturally counted by circulation – i.e. sold copies or web hits provided by Alexa[5]) and (b) geographic population density, the authors invented a measure of newspapers' influence in different locations throughout the country. Each "sphere of influence" has a radius linearly proportional to newspapers' circulation. Near coasts, the influence shape is not spherical but asymmetric, taking into account the fact that there is no population off shore.

[5] Alexa is a California-based subsidiary company of Amazon.com which provides commercial web traffic data. Its website is available at www.alexa.com

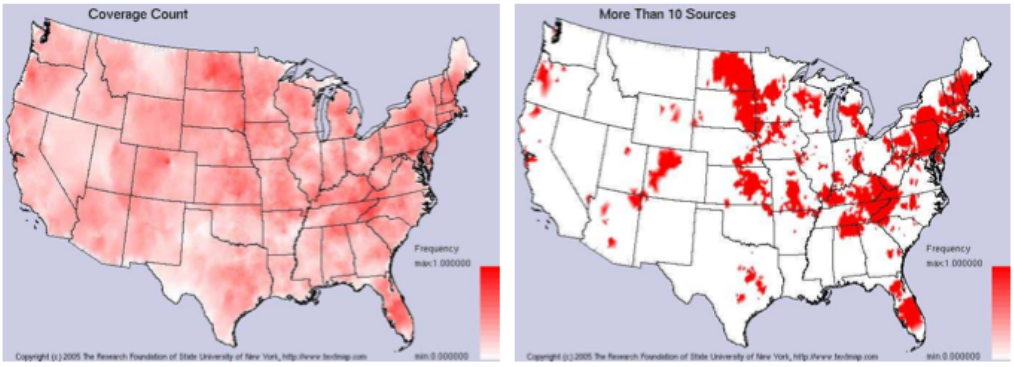


Figure 1 The number of different news sources influencing each U.S. city (left), and the number of cities influenced by more than ten sources (right). Figure taken from Mehler et al. (2006)

The *heat* associated to an entity *e* in a location *s* is given by the relative frequency of reference of *e* in each of the newspapers that have influence over *s*.

The influence of a source on a location can be heat mapped, in order to understand where different sources – e.g. the *New York Times* or the *Ithaca Times* – influence the most. The authors propose a function that takes in account the distance of the location from the source location. If the distance is longer than the radius, the influence of the source on such location is *zero*.

The resulting heat maps identify hot topics throughout the country and provide a view on how media coverage does not follow a uniform spatial distribution. Along with the paper, some examples are provided to show how an entity can be widely covered somewhere and completely ignored anywhere else.

The weak side of Mehler *et al* (2006) – which is due to its old age – is the importance they give to offline newspapers circulation in a way that does not consider: (a) the fact that online news are different from offline news – even when they are from the same newspaper group, as online editorial boards are often different from offline ones – so it does not make much sense to use the classic circulation measure to calculate the sphere of influence of newspapers; (b) the use of social networks to spread news, nowadays essential as explained by Kwak (2010).

Thus, something similar to what the Guardian Datablog (2012) proposed would be better, as their method tracks how articles are shared on social networks by combining geo-located tweets and IP-to-location services.

Another interesting work about Sentiment analysis for news comes from Bautin *et al* (2008). The researchers investigate how news sentiment analysis performs on cross-language news, by combining machine translation and Lydia sentiment analysis system (Lloyd *et al*, 2005). In particular, the authors analyse how international news talk about common named entities (such as countries or politicians) by automatically translating news from 8 languages (Chinese, Arabic and Korean included) to English and then inputting such translated news to Lydia. As there is no golden standard for international sentiment analysis, the authors based their evaluation on a measure of *consistency*, that is, to what extent entities-related sentiment scores (e.g. what polarity is associated to the entity *Iran*) differ from one language to another. They conclude that the significant correlation between English news entity polarity scores and machine-translated news entity polarity scores means that sentiment analysis tools (in particular, Lydia) could be adopted to make international sentiment analysis without losing performance, by including a machine-translation step in their pipeline (Lloyd, 2005). But their methodology contains unclear points: the authors do not explain how they assumed that two or more articles on the same day from different countries report the same facts. Apparently, instead, they only considered the entity frequency conservation as a clue for equivalence (reporting that only 19 entities were revealed in all languages, 18 of which representing geographical locations). Plus, the authors noticed that named entities can be referred by different ways in the various languages, but in order to produce *canonical names* for entities they only removed stop words such as articles, ignoring, for instance, shortened names. Finally, the authors provided a naïve conclusion by conjecturing that a real (i.e. *absolute*) entity polarity score might exist throughout all cultures. Such conclusion has to be rejected as facts are perceived and judged differently by different cultures (and sub-cultures), meaning that the correlation is not language dependant (as they stated), but source dependant. For instance, a political reform in a country can be perceived positively from some sources and negatively from others (even if speaking the same language).

2.2 Subjectivity and Emotion

To define what a sentiment or opinion is it could be necessary to recall many fields of knowledge, namely, psychology, neurosciences and even philosophy. Thus, when talking about opinion mining and sentiment analysis it is necessary to reduce the scope. Two specific concepts are related to our task, *subjectivity* and *emotion*, as described by Liu (2012).

An *objective sentence* presents some factual information about the world, while a *subjective sentence* expresses some personal feelings, views, or beliefs. (Liu, 2012:27)

An example of objective sentence is "Obama is the new president of the United States". An example of subjective sentence might be "I enjoyed the film Django Unchained". Subjective expressions emerge in several forms, for instance, opinions, desires, beliefs, speculations and others. However, not all the subjective sentences express sentiments, and not all the objective sentences do not carry or imply emotions. For instance, "I think she went fishing." is a subjective non-opinionated sentence, whilst "The yellow fever vaccine did not work properly and the disease spread across the Country."

Emotions are our subjective feelings and thoughts. (Liu, 2012:28)

Scientists have categorized people's emotions into some categories. However, there is still not a set of agreed basic emotions among researchers. Based on (Parrott, 2001) introduces six primary emotions, i.e., love, joy, surprise, anger, sadness, and fear, which can be sub-divided into many secondary and tertiary classes. According

to Liu “the strength of a sentiment or opinion is typically linked to the intensity of certain emotions” and “opinions that we study in sentiment analysis are mostly evaluations that can be “broadly categorized into two types: rational evaluations and emotional evaluations” (Chaudhuri, 2006).

Rational evaluations come “from rational reasoning, tangible beliefs, and utilitarian attitudes.” (Liu, 2012). For example, the following sentence express rational evaluations: «this article is well written». On the other hand, emotional evaluation comes “from non-tangible and emotional responses to entities which go deep into people's state of mind” (ibid.). An example of emotional evaluation follows: «I can't stand civilians are dying in Afghanistan». To make use of these two types of evaluations in practice, Liu (2012) designs five sentiment ratings: emotional negative (-2), rational negative (-1), neutral (0), rational positive (+1), and emotional positive (+2). Neutral often means that no opinion or sentiment is expressed.

2.3 Introduction to SentiStrength

Thelwall et al. (2010, 2012) proposed SentiStrength, a tool for sentimental analysis designed for short informal English text. The authors design their tool in order to detect positive and negative polarities of short, informal, ideally domain-independent, English text. SentiStrength includes both supervised and unsupervised approaches.

The supervised approach relies upon on social web data training sets and performs typically better for context-dependant texts that contain *indirect affective words* or idiosyncrasies; the unsupervised approach is based on a set of lexical rules as well as on a list of emoticons and other pragmatic clues. Both approaches return as output a pair of numbers (x, y), such as $0 < x, y \leq 5$, indicating respectively the positive and negative sentiments conveyed by the input sentence. As its main goal is viability more than performance, SentiStrength works well and provides further evidence of the robustness and versatility of unsupervised sentiment analysis methods based on lexicon and explicit rules. Problems arise when dealing with sarcastic utterances, while domain-specific tasks can be well performed by using supervised algorithms (if human-coded data is available).

In our experimental framework we use SentiStrength configured to retrieve from an input text its "trinary polarity", meaning that the algorithm returns the input's polarity making distinction similar to (Liu, 2012): in SentiStrength the output can be generally negative (-1), generally positive (+1) and neutral (0). SentiStrength considers both rational and emotional evaluations as sources of information to detect sentiment polarities – considering them on the same level.

In our research we use the unsupervised SentiStrength. Below, we report a list of SentiStrength's key features as reported in (Thelwall, 2011):

- A sentiment word list with human polarity and strength judgements. Some words include Kleene star stemming (e.g., *ador**).

- A spelling correction algorithm deletes repeated letters in a word when the letters are more frequently repeated than normal for English or, if a word is not found in an English dictionary, when deleting repeated letters creates a dictionary word (e.g., *hellp* -> *help*).
- A booster word list is used to strengthen or weaken the emotion of following sentiment words.
- An idiom list is used to identify the sentiment of a few common phrases. This overrides individual sentiment word strengths.
- A negating word list is used to invert following emotion words (skipping any intervening booster words).
- At least two repeated letters added to words give a strength boost sentiment words by 1. For instance *haaaappy* is more positive than *happy*. Neutral words are given positive sentiment strength of 2 instead.
- An emoticon list with polarities is used to identify additional sentiment.
- Sentences with exclamation marks have a minimum positive strength of 2, unless negative.
- Repeated punctuation with one or more exclamation marks boost the strength of the immediately preceding sentiment word by 1.

A second version of SentiStrength implemented some improvements to better detect the negative sentiments. In particular, the list of negative terms was extended from 693 to 2310 words; negation of negative terms makes them neutral rather then positive; the idiom list was extended with phrases indicating word senses for common sentiments (Thelwall, 2011).

SentiStrength was tested with six datasets coming from users' comments, forum discussions and social networks, namely: BBC Forum posts, Digg posts, MySpace comments, Runner World forum posts, YouTube comments and Twitter posts and a combination of all the datasets. In the Twitter dataset were included tweets of any type. More specifically, the dataset included 4218 tweets consisting on

average of 15.35 words and 94.55 chars each. The accuracy reached by the unsupervised version of SentiStrength within the Twitter dataset is 59.2% for positive comments and 66.1% for negative comments.

2.4 The connections between Twitter and the online news

Kwak et al. (2010) analysed Twitter from two points of view. First, they dissected Twitter's network nature employing Social Network Analysis methods. Second, they considered the impact of news on Twitter and the life and nature of Twitter trends (*i.e.*, tweets or topics which get into the social hype and ignite long-lasting discussions).

In order to address these two analytical approaches, they took a snapshot of the complete Twitter space – the so-called *Twittersphere*. The dataset was crawled in 2009 (from June to September). This could lead to consider such research obsolete, as Twitter's size has been quickly increasing in terms of size. Anyway, the dataset is big enough to be considered it as a significant sample of the current network. It includes: 41.7 million of user profiles; 106 millions of tweets. A positive side of their data collection is their spam filter, based on Clean Tweets – which allowed to filter out tweets from users who had been on Twitter for less than one day and tweets containing more than 2 trending topics; finally, the dataset tracked 4.262 trending topics and their tweets.

Several subtasks were involved to draw an overview of Twitter's structure included: basic analysis (number of followers and followings), reciprocity analysis, degree of separation, homophily investigation. In particular, throughout Twitter 77.9% of connections resulted to be not reciprocal, that is much lower than other social networks such as *Flickr,* which counts 68% undirected edges, as showed by Cha et al. (2009). Moreover, 67.6% of Twitter users were not followed back by any of those they follow. Such lack of reciprocity led the authors to reflect on the *raison d'être* of Twitter, labelling it as a news media rather than a social network in the classic meaning.

However, despite the lack of reciprocity, the average *degree of separation* is not as high as expected. In fact, there have been measured 4.12 hops on average to reach one node to any other in the Twitter direct graph (in particular, throughout its

giant component) – an even shorter average path than the one in well-known social networks as, for instance, MSN (Leskovec et al., 2008) in which 90% of users have degrees of separation equal to 7.8. Such trait makes Twitter different from classic social networks.

It happens with *homophily* and *followers distribution*. Homophily emerges only between reciprocally followed users, which are similar in terms of geographic location, popularity and – as showed by Weng et al. (2010) – interests. Followers' distribution does not follow a power law. *En masse*, Twitter has a particular network structure that includes traits typical of classic social networks, and traits more common among news media systems.

The second part of Kwak *et al.* takes into account Twitter trends and compares them to other media trends. By comparing Twitter trends to CNN Headline, half of them overlap in terms of topics. But if this look as a good observation, Kwak *et al.* seem not to consider Google Trends in a proper way. In fact, to match Twitter Trends to Google Trends, the authors implement a method based on *Longest Common Substring*, that is: one Google Trends is the same of one Twitter Trend if the length of the longest common substring is more than 70% of either strings. This way, only 3.6% of trends matched (126 out of 3497 unique topics), creating an ambiguous and weak evidence of trends mismatch. Basing the comparison on, for instance, semantic distance could have led to several matches more.

Measurements about nature and lifetime of trends were performed. For instance, it emerged the existence of core sub communities of users keeping the discussion high on the long term, although most of trends have one-week of lifetime and only 7% of trends persist longer than 10 days. This gives a view on how long Twitter audience could focus on a specific topic, before letting it decay.

No information about trends' contents was provided. Cheong (2009) tracked trends' contents and draws a pie chart, clustering trends in categories. In other terms, it provided an insight into Twitter's *Zeitgeist*. Mainly, trends speak about four categories: entertainment, sport, tech and meme – which all together cover 79%.

Only 6% are included in activism (e.g. #iranelection) and 1% into culture: the same amount of trends that speak about *Twitter* itself.

Kwak *et al.* concluded their research examining the impact of tweets. By considering Twitter's social graph's structure, they stated that the source's number of followers does not influence the tweet's audience as long as it starts to get spread over Twitter. A strong conclusion based on two weak assumptions: (i) tweets would be read by all the user's followers and (ii) the number of retweets is non-dependant on user's followers. Moreover, they did not specify how many times a tweet should get retweeted in order to reach a specific number of users on average (that is, what is the critical mass needed to reach *x* users?).

Abel et al (2011a, 2011b), in the context of user modelling, propose a set of methods to retrieve tweets that mention or, more generally, refer to news articles. Such methods are aimed to mine the contents of mentioned news to infer and model users' personal interests. To link their 458,566 tweets to news articles two strategies were followed: 98,189 relations were retrieved from explicit URLs in the tweets. 360,377 relations were deduced by comparing the named entities – extracted using Named Entity Recognition methods provided by OpenCalais.com – that appeared in both articles and tweets, taking in account news' and tweets' temporal contexts. The former method is supposed to be accurate (since links are explicit), the latter has been evaluated in (Abel, 2011b, section 4) and achieved 70% of precision and 15% of coverage. The datasets are available for research purposes (TweetUM, 2011) and their conclusion underlined the importance of news in Twitter as a mean to investigate users (and readers) nature and behaviour.

An *et al* (2012) deepened the ties between newspapers and Twitter by proposing a visualization method to show media bias through the social network analysis of relevant newspapers' Twitter accounts. By 'media bias' the authors intended what is commonly known as *political leaning* or *slants*. The model they proposed attempted to map media sources along a dichotomous political slants spectrum (i.e., from left to right wing, via centre leaning). The model relied on the

intuition that the closer two news sources are the more their Twitter audiences' overlap; in other words, the more Twitter users co-subscribe to two news sources, the more politically close those two sources are. The criterion behind their intuition is the similarity principle, a well-known social networks' pattern that lets people with similar interests and attitudes cluster together. In order to quantify the distance between two sources, a closeness metrics was used. To compute the actual position of a media source on the line spectrum, the authors adopted a Global Network Positioning algorithm (Ng and Zhang, 2002) which, fixed two landmarks (i.e., two sources with specified coordinates), allowed to place the other news media sources on the space, at a distance derived from the closeness measure. The final visualization is showed in Illustration 2.

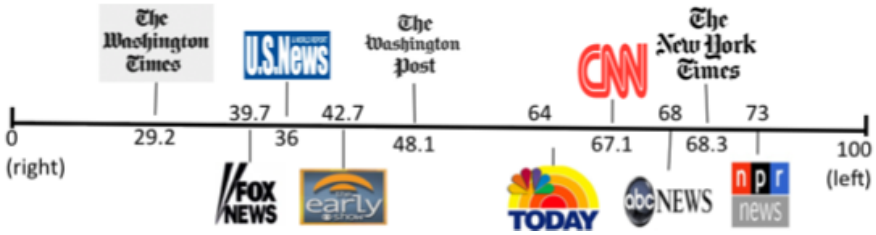


Figure 2 The output of An et al. model, taken from (An et al., 2012)

In order to evaluate their method, the researchers selected a gold standard: ADA, *Americans for Democratic Action* (Mylio et al, 2005). ADA is a well-known index assigning news sources a score from 0 to 100 indicating their political slants (being 0 far-left and 100 far-right), based on several quantities such as the number of citations of think-tanks and other politically leaned groups. By comparing their method to the gold standard, the authors detected a high correlation between the two lists. A weak point of their model is the choice of the two landmarks, directly taken from the ADA. Further works should focus on the automatic selection of landmarks, but the authors did not point out how that might be feasible.

An important issue related to media is information credibility. If Twitter can be considered a pseudo news media, it comes naturally to ask whether Twitter is a

credible source of information or not. This question is central in the research of Castillo *et al* (2011), who address the problem by designing automatic methods to analyse the credibility of a given set of tweets. To achieve this task the authors both develop and evaluate a set of algorithms that take into account the way tweets spread through Twitter social network. In fact, they show how propagation dynamics are a good credibility indicator. By comparing their automatic predictions to human-annotated assessments, the authors reach precision and recall of 70% and 80%, respectively.

Castillo *et al.*'s methods pivot on the intuition that hoax tweets follow patterns in terms of author's user profiles, topic, and sentiment polarity and propagation modality. Such patterns should be – in the researchers' opinion – enough to fill the gap left by virtuality in terms of clues that users have in real life, which asses the credibility of the information they are constantly exposed to. A definition of information credibility is provided by Fogg and Tseng (1999), which describe it as "a perceived quality composed of multiple dimensions", suggesting that credibility can be quantitatively modelled.

Their overall strategy is pivoted on two supervised algorithms: the first is aimed to identify newsworthy tweets. In other words, such method finds tweets conveying news events separating them from chat messages. The second, given a set of news-related tweets, attempts to distinguish credible ones from hoaxes. Training sets were made through the involvement of the Mechanical Turk, submitting 383 *topics* – *i.e.* sets of an average of 100 tweets each – and requesting 7 human evaluators to classify each topic. Once manually gathered a news-related tweets set, a series of features were extracted in order to create a feature set useful to assess newsworthy tweets. Features can be divided in four categories: (a) message-based features, such as length, number of URLs, positive and negative sentiments, number of exclamation and question marks; (b) user-based features, that look into tweets' authors' characteristics; (c) topic-based features, that consider topic's aggregate measurements, such as average length and number of distinct hashtags; (d)

propagation-based features, that take into account the propagation tree, for example counting the its max and average depth.

These features were used first to classify newsworthy tweets and second, among those, to identify credible tweets. For the first task the authors reached good results: 92% F-score to classify newsworthy tweets. The most relevant features characterising newsworthy tweets were: URL inclusion (probably because those tweets often include news-links) and deep propagation trees.

For the second task, 86% instances were correctly classified (but F-score is not provided). Both tasks performed the best using J48 decision trees. Interesting enough is what came out from the feature-level analysis. The authors observed that text-based and author-based features are not powerful enough to distinguish credible and hoax news-tweets. On the contrary, propagation-based features plus the fraction of retweets, the total number of tweets and the fraction of tweets that contain the most popular URL, hashtags, mentions or author are shown to be the most relevant feature for assessing credibility. In general, non credible news-related messages look like having common propagation patterns; concerning the social network structure, it is remarkable that Twitter network works like a “social filter”, that is, users with longer experience propagate – mostly – credible news, improving both their reputation and the information circulating within the whole community.

What Castillo *et al* did not take in account is the semantics of tweets. It could be essential, for instance, to see what the typical contents of hoaxes are; or observing whether particular named entities – or categories – recur more frequently than others; or to explore tweeted URLs' content, to see how fake news are structured in a wider space such an online news articles. However, the direction researchers plan to pursue in the *future works* section is quite interesting: explore the features of hoax and credible news-related tweets without diving into the tree but focusing on the surface level of the network graph.

2.5 Data-driven journalism: context and definition

In the very last few years, parallel to the growth of the so-called Information Age, particularly due to the spread of the Internet, the concept of «Data-driven journalism» (or, more concisely, «Data journalism») became popular among news media. According to Bounegru *et al* (2012), a unique definition of 'data journalism' is not possible due to ambiguity of both terms: 'journalism' and 'data', which encompass several meanings themselves. For this reasons, the authors propose a definition by contrasting the differences between data journalism and the rest of journalism, pointing out new opportunities blooming by combining "the traditional 'nose for news' and ability to tell a compelling story, with the sheer scale and range of digital information now available." (Bounegru *et al*, 2012). This suggests a methodological approach that puts the data in the centre of the news development process,. However, such definition still does not dive into the epistemology of the idea. To understand what data journalism is, it is useful to go back to its roots, when data journalism had two other names: precision journalism and computer-assisted reporting (CAR), in the 1970s and 1950s, respectively.

Precision journalism was firstly introduced by Dennis (1974). He had the idea of applying social science research methods to journalism. Such genre arose as direct reaction to the so-called "narrative journalism", another current of journalism that involved fiction-like techniques such as internal monologue, detailed character development and scene settings to tell real and newsworthy stories. The most popular representatives of narrative journalism have been Gay Telese, Truman Capote and Tom Wolfe, who are considered the founders of narrative journalism, also known as "creative nonfiction", another way to name this genre aimed to tell well-documented stories holding the attention of the reader through narrative devices.

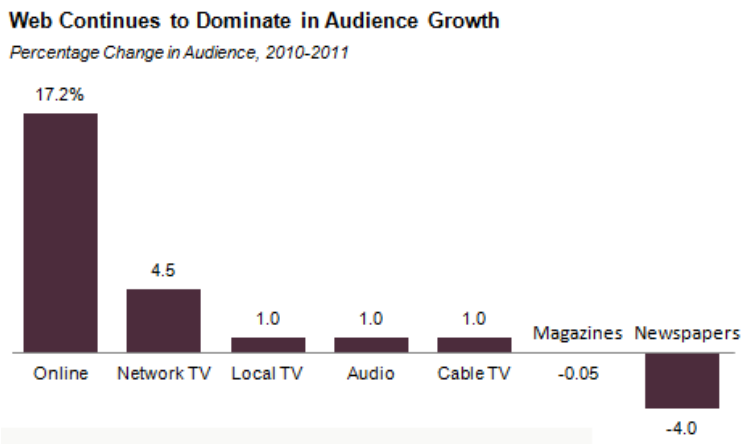
According to Meyer (2011) precision journalists started to practise it motivated by the feeling that narrative journalists "are subjective to a degree that disturbs conventional journalists and horrifies precision journalists. In essence, all the other

new journalists push reporting toward art. Precision journalists push it toward *science.*" Not by chance in France, the definition "precision journalism" has been translated as *le journalisme scientifique*, making clearer the philosophy behind it. In Meyer's definition being precision journalists means "treating journalism as if it were a science, adopting scientific method, scientific objectivity, and scientific ideals to the entire process of mass communication" (Meyer, 1991).

Proceeding backwards, computers started to be involved in journalism by a small community of US-based pioneers in the 1950s. The first milestone of Computer Assisted Reporting (CAR) was reached in 1952, when CBS first adopted it to predict the result of the presidential election. Even in this case, scientific methodology was a key-concept.

One of the cardinal characteristics of science is that its results can be properly verified. And so can be reported facts, if reported according to a scientific methodology. Data journalism, as with precision journalism, let stories emerge from a scientifically framed environment – emphasising on the role of data as a primary source. Data journalism is the natural evolution of precision journalism and CAR in times of data explosion. Meyer observed, as quoted in (Bounegru, 2012), that "when information was scarce, most of our efforts were devoted to hunting and gathering. Now that information is abundant, *processing* is more important". Data-driven journalism is then considered by Meyer as a double-layer process: first, "analysis to bring sense and structure out of the never-ending flow of data" (Bounegru, 2012); and second, data visualisation to make data-driven stories widely accessible.

Data journalism projects are achieved by adopting information science. The source data can be either numeric or text, structured or semi-structured or non-structured, machine- or human-generated. When coping with human-generated data, Natural Language Processing techniques are those to be adopted naturally. Potentially, the whole data-driven news production process might involve NLP: its methodology, algorithms and tools would allow to elaborate huge quantities of human-generated text, making sense out of it *automatically* (or almost, since the



human supervision can be essential and unavoidable). By applying NLP to social networks such as Twitter – where users post news directly from the scene, making then "citizen journalism" – crowdsourced stories can be revealed. Reporters, by combining language technologies and their experience, would be able to refine and select stories, for instance determining credible and newsworthy tweets (Castillo *et al*, 2011) but also searching for sources and stories, mining Twitter using advanced information retrieval systems involving NLP modules, to address tasks such as, for instance, NER, sentiment analysis, topic and event detection, information extraction, machine translation, text summarization and semantic analysis.

It is natural to consider information technology as the cause of the so-called *information overload*, but as Meyer (2011) noticed, technology can also be adopted to contain it. For instance, by settling NLP systems into editorial offices, data can be synthesized, aggregated and processed in order to distinguish sound (stories) from noise (non stories), improving journalism with a new, data-driven, science-oriented methodology.

The business of media industry started changing drastically since the Internet began to spread among people and societies. Last data released by the Pew Research Center reveal that decline of print circulation and ad revenues, as "in 2011, losses in print advertising dollars outpaced gains in digital revenue by a factor of roughly 10 to 1, a ratio even worse than in 2010. When circulation and advertising revenue are combined, the newspaper industry has shrunk 43% since 2000." (Pew Research Center, 2012). Elements such as online and mobile news growth, decline of

traditional media and necessity of new products, constitute rich soil to let data journalism increase its use among journalists and spread among users.

The actual value of data journalism has been confirmed by its recent growth rates. The Pulitzer Prizes of 2009, 2010, 2011 and 2012 have been regularly given to online newspapers developing, alongside traditional journalism, data-driven applications.

The Pulitzer prize received by the St. Petersburg Times' Politifact in 2009 has been pointed out as a turning point by the Pulitzer Aron Pilhofer, New York Times, who commented: "The Pulitzer Prize going to PolitiFact – the first Web project to be so honored – is a watershed moment for journalism, I believe, much like *The Color of Money* which 20 years ago was the first Pulitzer awarded to a project that relied heavily on statistics and data analysis, what has come to be called 'CAR'. Two decades from now, we may very well refer to some significant event as a *PolitiFact moment*." (Pilhofer, 2009). But two decades are probably too long considering how fast media are running after technology. The 2010 and 2011 Pulitzer Prizes (Investigative and National reporting sections, respectively) have been awarded by ProPublica, an American online investigative news organization, who made large use of data and technology. In 2011, the awarded ProPublica's stories were only published online, not in print, becoming a milestone for online media. In 2011, also, the Herald-Tribune (2011) has been awarded by a Pulitzer for a database journalism project involving insurance policies in Florida, showing an interactive map. This enhanced the idea of transforming online news organization into public service portals.

In May 2012 the International Data Journalism Awards has been organized. The competition is the first contest for data journalists, founded by the Global Editors Network, the European Journalism Center and Google in order to witness and recognize the relevant value of data journalism. Six prizes have been given out for three categories, both awarded at national/international and local/regional levels:

Data-driven investigations; Data visualisation & storytelling; and Data-driven applications (mobile or web). Winners included the leading Guardian Datablog (2012), for their work on rumours during England's Summer Riots (Procter, 2011) – that also used SentiStrength to help gauge twitter's opinions about the riots and during the riots –, as well as emerging realities such as the Northern Irish online news-site *The Detail* (2012), which mapped ambulance service response times in the six Irish counties of Ulster. In 2012, also Italian newspapers started publishing data journalistic works, as for example "Patrie Galere" (lit. national prisons), an interactive map of deaths of prisoners in Italy in 2002–2012, published by *Il Fatto Quotidiano*[6] and featured on *The Guardian Datablog*.

[6] An English version of the "Patrie Galere" project can be found at http://www.ilfattoquotidiano.it/patrie-galere-deaths-italian-prisons-since-2002-2012/ and an introduction to the project can be found on *The Guardian Datablog* http://www.guardian.co.uk/news/datablog/2012/may/23/italian-prisoners-deaths

Chapter 3
Methodology

We approach the news-oriented Twitter sentiment analysis by answering our research questions. To do so, we follow a series of steps – involving different technologies that are used in the current data journalism scene.

First, data collection: we have to build a corpus including the class of tweets we want to analyse (the news-linking tweets). Secondly, it is necessary to pre-process our data to (a) filter out tweets that have been wrongly included in the corpus, and (b) enrich them, namely downloading the title of the linked news, and finally (c) refine them, making final adjustments to the data to prepare it to be inputted to SentiStrength.

Once the corpus is ready, it was submitted to SentiStrength. In this phase the polarities of the tweets are detected, both in the complete tweet (including title and comment) and in the comment-only version. For every tweet–format we have the polarity (-1 for negative, 0 for neutral, +1 for positive).

To evaluate the results provided by SentiStrength, we created a gold standard, which can be used to evaluate the performance of such tool in the context of news-linking tweets and draw the first conclusions. The gold standard is a sample of the corpus, which was randomly selected and manually annotated. An inter-annotator agreement was, then, conducted to have reliable estimations of the polarities manually assigned.

3.1 Data Collection

In order to set up our experimental environment, we gathered a set of news-linking tweets from the Twitter API[7]. In order to download the tweets a scraper was set to query the Twitter API on a daily basis; the scraper was based on ScraperWiki[8] and had been crawling for a period of six months. Only tweets in English were collected. ReTweets (RTs) were filtered out by the query. In order to include only news-linking tweets, we chose *The Guardian* newspaper's website. This way, it was possible to download tweets via the Twitter API by selecting those that contained the Guardian's web domain in them, in particular his twitter-shortened version (*gu.com*). An example of gu.com shortened URL is "http://gu.com/p/3f266/tf".

[7] The documentation of Twitter API can be found at https://dev.twitter.com
[8] ScraperWiki.com is a Liverpool based start-up providing web-scraping services. Scrapers can be coded in Python, Ruby and PHP.

The description of the parameters of our Twitter API query follows.

N.	Twitter API parameter	Description
1	QUERY = 'gu.com -RT'	Tweets including The Guardian's domain *gu.com* without RTs
2	RESULTS_PER_PAGE = '1000'	Each page will include 1000 tweets
3	LANGUAGE = 'en'	Tweets in English
4	NUM_PAGES = 100	The first 100 pages of Twitter search

Table 1 Parameters of our Twitter API query

Parameters 2 and 4 allowed us to download the maximum number of tweets allowed by the Twitter API Term and Conditions.

The final set of tweets counted 24972 instances, all recorded on a remote SQL database and available to be downloaded in open data format (e.g., CSV) and preprocessed.

3.2 Pre-processing

In order to create a sample of data to be manually annotated and thus used to test SentiStrength's performances, a subset of 7215 tweets was randomly selected. The precise news' title was downloaded for each of these tweets' links. This step was achieved by querying the Guardian Open Platform's API[9].

The Open Platform is a "suite of services for developing digital products and applications with the Guardian" and allows developers to download datasets from the Guardian's website. In our case, we queried the Guardian API to associate to every tweet the news title. The choice to select a subset of 7215 tweets was driven by the restrictions imposed to the queries to the Guardian API, that only permits a limited number of hits per day.

The first table of our news-linking tweets corpus had the following schemata.

ID	Tweet	URL	News Title	Other Tweet Info

An example of an entry follows:

3215839286 09751040	An intense comment on the war in Serbia, 15 years later http://gu.com/p/2t4hq/tw	http://gu.com/p/2t 4hq/tw	War Child and the Bosnian war 15 years on	…

The «Other Tweet Info» included: author, timestamp, geolocation, mentions, and other tweets' information that can be retrieved from the Twitter API.

[9] A complete introduction of the Guardian Open Platform can be found at http://www.guardian.co.uk/open-platform

3.3 Corpus description

From the collected 7215 tweets, 3257 included the news-link, the exact title of the news – as reported by the Guardian API – and another sets of words (excluding the "via @guardian"). This means that those tweets include most probably a comment; the remaining 3958 tweets did not contain any of those elements and, thus, were excluded. The average length of the complete and clean tweets is, respectively, 116 and 31 characters.

From every tweet with title, news link and comment, the comment was extracted and recorded apart. At this point SentiStrength was run on both the complete tweets and the comment parts. The output of SentiStrength is set to be the *trinary polarity*, i.e., 1 for positive, 0 for neutral and -1 for negative tweets. The two outputs for every record were saved, one output for the complete tweet (news title + comment), one for the clean tweet (only comment).

The final table of our news-linking tweets corpus had the following schemata.

ID	Tweet	News Title	Comment	Timestamp	Tweet's Polarity	Comment's Polarity	URL	Author	Other Tweets Info

The polarity of the 3257 tweets with title and comments is distributed as follows:

Polarity	Original Tweets	Original Tweets %	Clean Tweets	Clean Tweets %
-1	1297	39.82	521	16.00
0	1003	30.80	1749	53.70
+1	957	29.38	987	30.30
Total	3257	100.00	3257	100.00

Table 2 Polarity distribution of the tweets sample. The tweets are submitted to SentiStrength in both their complete and clean forms.

It can be observed that the clean tweets result to be apparently more neutral and less negative than the complete forms (made negative by the titles). The number of positive clean tweets remains, on the other hand, almost the same.

3.4 Discordant tweets

We define “discordant” those tweets that once processed by SentiStrength return two different outputs, depending on the inclusion or exclusion of the news title. Of the 3257 tweets that have been submitted to SentiStrength as input, 1190 were discordant (36.5%). In other words, those tweets could let SentiStrength perform better (or worse) when the noisy (or informative) part of the title is kept apart. An example of discordant tweet follows:

From last year, but still interesting… Ruby Wax: depression, me and you http://gu.com/p/343fz/ via @guardian

In this case, the title of the news is «Ruby Wax: depression, me and you» and the comment is «from last year, but still interesting…» When the whole tweet is given as input to SentiStrength it returns a negative overall polarity (due to the presence of the term «depression» that strongly influences the evaluation), whilst when the input is only the «from last year, but still interesting…» polarity is positive (in this case, due to the presence of the term «interesting»).

3.5 Annotation

In order to build a gold standard, we manually annotated a sample of 318 (approx. 10% of the entire collection). The annotation task consisted in reading the whole tweet (only the original text) and determining the polarity (positive, negative or neutral) of the opinion expressed in the user's comment. The annotator had the opportunity to see the news' title in a column apart when it was not clearly distinguishable in the tweet. To measure the reliability of the polarity annotation, we conducted an inter-annotator agreement trial, involving two annotators. Of these tweets, a sample of 99 has been selected[10] and submitted to the second annotator. The first annotator was an Italian masters student in Natural Language Processing (myself). The second annotator was a Portuguese professor in linguistics already familiar with sentiment analysis related annotation tasks. An inter-annotator agreement was made to have a more reliable estimation of the polarities.

The results of our inter-annotation process follow:

% Agreement	Krippendorff's Alpha	N Agreements	N Disagreements	N Cases
96.0	0.936	95	4	99

A percentage agreement of 96% and Krippendorff's Alpha of 0.936 have been considered good enough for our purposes and enforced the reliability of the complete annotation made by only one annotator. We used Krippendorff's alpha as defined in Krippendorff (2004).

[10] The sample was originally made of 100 tweets. They became 99 after one of the tweets was detected to be bogus (i.e., without an English comment) and excluded.

The polarity distribution of the complete annotated set (318 news-linking tweets) follows.

	# Tweets	Percentage (%)
+1	113	35.53
0	124	38.99
–1	81	25.47
Total	318	100.00

Table 3 Polarity distribution of the complete annotated set.

Chapter 4
Experiments and Results

To have a first view on the results of our experiments, we have counted the number of "concordant tweets" by comparing the annotated tweets with SentiStrength's output both with complete (title + comment) and clean (only comment) tweets. This measure is what is also known as *recall* (i.e., the ratio between the number of correctly labelled tweets with polarity *p* and the number of tweets actually with polarity *p*).

The results are showed in the following table:

Polarity \ Type	Total annotated tweets	Concordant complete tweets (Recall)	Concordant complete tweets (Recall) (%)	Concordant clean tweets (Recall)	Concordant clean tweets (Recall) (%)	Both non-concordant	Both non-concordant (%)
+1	113	19	16.81	74	**65.49**	20	17.70
0	124	4	3.23	109	**87.90**	11	8.87
–1	81	58	**71.60**	12	14.81	11	13.58
Total	318	81	100.00	195	100.00	42	100.00

Table 4 Results of our test with the annotated test.

It is also worth noting how 13.21% of the tweets are wrongly predicted both when submitted as clean and complete. Theoretically, this sample of tweets provides elements to evaluate what causes a wrong evaluation and opens a door for future research. On the other hand, this result suggests that 87.79% (excluding the 13.58% from the total) of tweets can be correctly predicted if it is made the distinction (see the pseudo-code in the end of the chapter).

The confusion matrix for the complete tweets follows:

Complete tweets (title + comment)			
Actual \ Predicted	**+1**	**0**	**–1**
+1	19	7	87
0	34	4	86
–1	9	14	58

Table 5 Confusion matrix of the results with complete tweets

The confusion matrix for the clean tweets follows:

Clean tweets (only comment)			
Actual \ Predicted	**+1**	**0**	**–1**
+1	74	39	0
0	8	109	7
–1	27	42	12

Table 6 Confusion matrix of the results of the test with clean tweets

From these confusion matrixes it is possible to calculate Precision, Recall and f-Measure. In particular, *precision* shows the ratio between the number of correctly labelled tweets with polarity *p* and the number of tweets labelled with *p*. *Recall* indicates the ratio between the number of correctly labelled tweets with polarity *p* and the number of tweets actually with polarity *p*. *f-Measure*, finally, can be interpreted as a weighted average of the precision and recall (i.e., the harmonic mean of precision and recall).

The formula of f-Measure follows:

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

	Complete tweets (title + comment)			Clean tweets (only comment)		
Actual \ Predicted	**Precision**	**Recall**	**f-Measure**	**Precision**	**Recall**	**f-Measure**
+1	0.31	0.17	0.22	0.68	0.65	**0.66**
0	0.16	0.03	0.05	0.57	0.88	**0.69**
–1	0.25	0.72	**0.37**	0.63	0.15	0.24
Average	0.31	0.24	0.21	**0.56**	**0.63**	**0.53**

Table 7 Precision, recall and f-Measure of the tests with complete and clean tweets.

As it can be observed in the results tables, from an overall perspective the cleaned tweets (i.e., without title) allow SentiStrength to reach a better performance both in terms of Precision, Recall and f-Measure. Specifically, the average f-Measure with clean tweets is higher (+0.32); while recall increases of +0.59 and precision of +0.25.

If we consider the f-Measure, it is relevantly better for the clean tweets when dealing with positive and neutral polarities (0.66 instead of 0.22, and 0.69 instead of 0.05) – while with negative comments, complete tweets work better (f-Measure of 0.37 instead of 0.24).

Interestingly, with tweets with negative polarity the recall is higher with complete tweets (71.6% of recall). This probably occurs for two reasons: because (a) news titles often have a strong negative polarity and the user « exploits » that to carry his negative emotion and (b) the majority of titles in our sample of 318 tweets – when submitted to SentiStrength – return a negative polarity (72.6% are negative, 20.8% are positive, 6% are neutral). These two elements boost the recall for negative tweets when evaluated from the complete version.

Beyond the general improvements indicated by Thelwall (2012), as, for instance irony detection, word sense disambiguation or a more general extension of the lexicon, in a hypothetical framework aimed to apply SentiStrength to news-linking tweets our results can be taken in account and – to obtain better performance and more reliable outputs – act as follows:

```
Input: tweet t
Output: trinary polarity (1, 0 or -1)

  1. t = pre-process(t)

  2. p = SentiStrength (t)

  3. If the p is negative:
     return p

  4. Else:
        t1 = clean_tweet(t)
        p = SentiStrength(t1)
        return p
```

Figure 3 Pseudo-code of the proceedure to use with SentiStrength to handle news-linking tweets.

In the pseudo-code above, it is showed a possible procedure to be used as an improvement of SentiStrength in the particular context of news-linking tweets. This improvement does not touch SentiStrength's algorithm but only manages to provide a better input and does not create problems of efficiency to the algorithm.

Chapter 5

Conclusions

In our work we have explored the connections between social networks, sentiment analysis and news media. We chose three entities, namely, Twitter, SentiStrength and The Guardian and we combined them. In particular, we built a corpus of news-linking tweets (i.e., tweets that include links to news articles, titles and comments) and tested SentiStrength's performance when dealing with them.

We concluded that generally SentiStrength performs significantly better if it excludes the titles from the tweets. This happens because the titles contain usually strongly opinionated words and influence SentiStrength's output.

Besides the experimental results, the corpus we built to answer our research questions allowed us to lay the foundation of a wider platform for sentiment analysis applied to the news sphere. By analysing the data we have gather it will be possible to focus on readers' reactions to news and visualize it in various forms and inspiring new data journalistic applications. In fact, such platform automatizes a series of steps to play with Twitter data, from data collection to sentiment analysis, and opens the doors to applications that take advantage of the relationships and connections between Twitter, online-news and opinion mining.

References

Abel, F., Gao, Q., Houben G., and Tao, K. 2011. – Analysing user modelling on twitter for personalized news recommendations. In *Proceedings of the 19th international conference on User modelling, adaption, and personalization (UMAP'11)*. Springer-Verlag, Berlin, Heidelberg, 1-12.

Abel, F., Gao, Q., Houben, G.J., Tao, K. 2011. Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. Technical report, submitted to: *Extended Semantic Web Conference (ESWC)*, Heraklion, Greece (2011).

Agrawal R., Rajagopalan S., Srikant, R. and Xu, Y. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of WWW*, pages 529–535, 2003.

An, J., Meeyoung, C., Krishna, P. G., Crowcroft, J., Quercia, D. 2012. Visualizing the media landscape through Twitter in *SocMedNews 2012 AAAI ICWSM Workshop*.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva J. 2010. Sentiment analysis in the news. *Proceedings of LREC.* Volume 10.

Bautin, M., Vijayarenu, L. and Skiena, S. 2008. International Sentiment Analysis for News and Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'08)*.

Bounegru, L., Chambers, L., Gray, J. eds. 2012. *The Data Journalism Handbook.* Princeton, NJ: Princeton University Press.

Bostock, M. 2012. D3.js is a small, free JavaScript library for manipulating documents based on data. Available at: http://mbostock.github.com/d3/ [Accessed 18 March 2012]

Boyd, D.; Golder, S.; Lotan, G. 2010. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," *System Sciences (HICSS), 43rd Hawaii International Conference on* , vol., no., pp.1-10, 5-8 Jan. 2010

Castillo C., Mendoza M., and Poblete B. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (WWW '11). ACM, New York, NY, USA, 675-684.

Cha, M., Mislove, A., Gummadi, K. P. 2009. A measurement-driven analysis of information propagation in the Flickr social network. In Proceedings of the 18th international conference on World Wide Web. Pages 721-730. ACM, 2009.

Cheong, M., 2009. What are you Tweeting about? A survey of Trending Topics within Twitter. *Search*, p.1-12.

Dove, G. and Jones, S. 2012. Narrative Visualization: Sharing Insights into Complex Data. Paper presented at the Interfaces and Human Computer Interaction (IHCI 2012), 21 - 23 Jul 2012, Lisbon, Portugal.

Godbole, N., Srinivasaiah, M., Skiena, S. 2007. Large-Scale Sentiment Analysis for News and Blogs, in *Proceedings of the International Conference on Weblogs and Social Media*. ICWSM 2007. Boulder, Colorado, USA.

Guardian News and Media. 2012. *The Guardian Datablog: How Bitly mapped Britain's news websites for us.* Posted by Simon Rogers on the 16th of May 2012. http://www.guardian.co.uk/news/datablog/2012/may/16/bitly-news-map-britain-data [Accessed 27 May 2012].

Guardian Datablog. 2012. *Reading the riots, investigating England's summer of disorder. How misinformation spread on Twitter during a time of crisis.* Available at: http://www.guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter. [Accessed 4 Jun 2012]

Herald-Tribune. 2011. *Florida's insurance industry investigations by Paige St. John.* Available at: http://projects.heraldtribune.com/insurancerisk/insuranceriskhome.html [Accessed 4 June 2012]

Kwak, H., Lee, C., Moon, S. 2010. "What is twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*. New York, NY, USA – ACM, 2010, pp. 591–600.

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Pages 219-250. Thousand Oaks, CA: Sage.

Lee, L. and Pang, B. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (January 2008), 1-135.

Leskovec, J. and Horvitz, E., 2008. Planetary-scale views on a large instant-messaging network. *Proceeding of the 17th International Conference on World Wide Web WWW 08*, p.915.

Liu, B. 2012. Sentiment Analysis and Opinion Mining (Introduction and Survey). Morgan & Claypool Publishers, May 2012 (167 pages).

Lloyd, L.; Kechagias, D.; and Skiena, S. (2005). Lydia: A system for large-scale news analysis. In Symposium of String Processing and Information Retrieval (SPIRE '05), 161–166.

Meyer, P. 1991. *The New Precision Journalism*. Rowman & Littlefield, 1991.

Milyo, J., and Groseclose, T. 2005. A measure of media bias. *The Quarterly Journal of Economics* 120(4):1191– 1237.

Meyer, P. 2011. *Precision Journalism and Narrative Journalism: Toward a Unified Field Theory*. The Nieman Foundation for Journalism at Harvard. Available at: http://www.nieman.harvard.edu/reports/article-online-exclusive/100044/Precision-Journalism-and-Narrative-Journalism-Toward-a-Unified-Field-Theory.aspx [accessed 1 June 2012]

Mullen, T. and Malouf, R. (2008). Taking sides: User classification for informal online political dis- course. *Internet Research*, 18:177–190.

Ng, T. S. E., and Zhang, H. 2002. Predicting internet net- work distance with coordinates-based approaches. *In Proceedings of the Infocom*.

Pew Research Center. 2012. The State of the News Media 2012, An Annual Report in American Journalism. Key findings. Available at: http://stateofthemedia.org/2012/overview-4/key-findings/. [accessed 4 June 2012]

Pilhofer, A. 2009. A PolitiFact Moment for Journalism. Available at: http://blog.morethanthis.net/2009/apr/22/links_for_20090_65/ [accessed 4 June 2012]

Procter, R. 2011. How 2.6m tweets were analysed to understand reaction to the riots. Available at: http://www.guardian.co.uk/uk/2011/dec/07/how-tweets-analysed-understand-riots [accessed 29 March 2013]

Schleimer, S., Wilkerson, D., Aiken, A. 2003. Winnowing: Local algorithms for document fingerprinting. *22th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems,* p. 76–85, San Diego, California, USA.

Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social Web. *Journal of the American Society for Information Science and Technology*, 63(1), 163-173.

The Detail. 2012. *How quickly did help arrive where you live?* By Kathryn Torney. Available at: http://www.thedetail.tv/issues/72/ambulance-response-times/how-quickly-did-help-arrive-where-you-live. [accessed 4 Jun 2012].

TweetUM – TU Delft, Web Information Systems. (2011). *TweetUM: Twitter-based User Modelling Framework.* Available at: http://wis.ewi.tudelft.nl/umap2011/#dataset. [Accessed 14 Mar 2012].

Weng, J., Lim, E.-peng & Jiang, J., (2010). TwitterRank: Finding Topic-sensitive Influential Twitterers. *New York*, Paper 504, p.261-270.