Chapter 8

Investigating Diversity in News Sources

Ilija Subašić and Bettina Berendt. Peddling or creating? Investigating the role of twitter in news reporting. In Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR'11). Springer-Verlag LCNS, Berlin, Heidelberg, 207-213. 2011.

Contributions as first author:

- (a) Related work section;
- (b) Similarity framework definition;
- (c) Conducting the case study;
- (d) Co-interpretation of the case study results.

8.1 Abstract

The widespread use of social media is regarded by many as the emergence of a new highway for information and news sharing promising a new information-driven “social revolution”. In this paper, we analyze how this idea transfers to the news reporting domain. To analyze the role of social media in news reporting, we ask whether citizen journalists tend to *create* news or *peddle* (re-report) existing content. We introduce a framework for exploring divergence between news sources by providing multiple views on corpora in comparison. The results of our case study comparing Twitter and other news sources suggest that a major role of Twitter authors consists of neither creating nor peddling, but extending them by *commenting* on news.

8.2 Introduction

On January 16, 2009, a US Airways airplane made an emergency landing on the Hudson river. First reports on this events were spread via social media web sites. Although the idea of “citizen journalism” was present much before this incident took place, it has provided a major boost to citizen journalism platforms, placing them shoulder to shoulder with “traditional” news outlets. By some [14], this new way of discovering news is hailed as a beginning of a “social revolution” driven by information sharing, promising stronger social action and making *vox populi* a more important factor in all spheres of life. However, some researchers [8, 6] have expressed doubts about such a social-media-led revolution. Transferring the same principles and contrasting standpoints to the news reporting domain, one could expect that social media either have a great potential for introducing and spreading new information, or alternatively serve solely as a channel for spreading the content produced by traditional media. Thus, is the (main) role of citizen journalists to *create* news or rather to *peddle* (re-report) existing content? In this paper, we aim to provide some insight into this question by defining a set of corpora-similarity measures on corpora created from Twitter and other news sources. The main idea of using corpora similarity is that higher similarity would suggest “peddling”, while lower would suggest originality and “creation”.

There has been substantial research into discovering the point of origin of a news story [7] and into the dynamics of content between news and blogs. We take a different approach: we start with news story already “discovered” and investigate whether social media provides a *different* reporting to traditional media. We start by collecting corpora containing documents on the same news story originating from different sources. Our goal is to analyze differences

between corpora by providing a multi-aspect view on similarity. Some news stories describe breaking events or spotlighted topics. These stories are often referred to as “breaking news”. We broaden our analysis and investigate whether during a breaking event reporting converges across sources.

The main contribution of this paper is a framework for comparing social media with traditional media that provides: (a) multiple-aspect corpora difference measures, (b) analysis of social vs. traditional media content, and (c) aggregation and visualization of news sources relations. We complement the framework (Section 8.4) by a case study (8.5).

8.3 Related work

Twitter research. Out of many areas of Twitter research, we focus on the ones related to news mining. [9, 16, 8] investigate user motivation behind twittering. All of these studies report on news sharing as one of the main motivations for Twitter use. Studies of the role of news medium in influence spread [16, 3] found that traditional news sources and celebrity-owned Twitter accounts were among the most influential posts. In contrast to these works, our objective is to detect the differences between news reports covering the same story on Twitter and other media.

Corpora and text similarity. Similarity between texts has been a long-standing topic in different fields producing a wide range of text similarity measures. [2] provides a valuable overview of different text similarity measures. Work in corpus linguistics [10] compares text similarity metrics on a corpus scale. This work introduces a χ^2 -test based model of corpus similarity and compares it with the probabilistic similarity measures perplexity [5] and mutual information [4]. Another family of probabilistic similarity measures, based on Kullback-Leibler (KL) divergence [11], has been widely used in different domains as a measure of text similarity. It is used for measuring similarity between queries and documents in information retrieval [12], for detecting plagiarism in Wikipedia articles [2], and for comparing traditional and Open Access medical journals content [15]. We adopt a KL-based approach to corpora similarity, but provide multiple perspectives on similarity by combining several aspects of the corpora.

8.4 Measures of corpora divergence

Notation. A corpus C_{source}^{story} is a set of documents covering the same news *story* collected from a single *source* (e.g. Twitter, AP) or a family of sources (e.g. blogs). A representation of a corpus C_{source}^{story} is a language model Θ_{source}^{story} , where the probability of a token t is denoted as $\Theta_{source}^{story}(t)$. We define token categories as: (1) plain words (pw) - words in a document; (2) headline words (hw) - words in headlines; (3) entity words (ew): words referring to a semantic entities (names, locations, companies, ...); and (4) sentiment words (sw): words expressing sentiment. $\Theta_{source|category}^{story}$ denotes the category language model (e.g. for a unigram model, $\Theta_{source|hw}^{story}$ are the probabilities of headline words).

Divergence measures. Among many different language models we choose the unigram model as it fits the writing style of tweets. Given two corpora from sources a and b covering a story x , we use a symmetrical variant of KL divergence, the Jensen-Shannon divergence (JS), between their language models Θ_a^x and Θ_b^x to measure their distance as:

$$JS(\Theta_a^x, \Theta_b^x) = \frac{1}{2}KL(\Theta_a^x, \Theta_m^x) + \frac{1}{2}KL(\Theta_b^x, \Theta_m^x); \quad (8.1)$$

where the probability of every t in Θ_m^x is the average probability of t in Θ_a^x and Θ_b^x . We define a set of measures differing by token categories.

Language divergence (LD). The first measure we define covers the entire content of the corpora. In other words, we build a language model using pw . The reason for this is to capture stylistic, terminological, and content differences of sources. We define language divergence (LD) of two sources a and b reports on a story x as:

$$LD_x^{a,b} = JS(\Theta_{a|pw}^x, \Theta_{b|pw}^x). \quad (8.2)$$

Due to many differences in the format and length between documents between corpora, using this measure mostly captures differences in writing styles and vocabulary between sources.

Headline divergence (HD). Headlines in traditional news summarize their articles; they are a standard unit of analysis in media studies. Tweets have no substructure and are at most 140 characters long, making them their own headlines ($hw = pw$). We define the headline divergence as:

$$HD_x^{a,b} = JS(\Theta_{a|hw}^x, \Theta_{b|hw}^x). \quad (8.3)$$

Using headlines to measure difference between reports in social and traditional media tackles problems of style and length used among sources. However, it still does not take into account the semantic difference between reports.

Named-entity divergence (ND). News stories revolve around different subjects, places, and organizations they describe or “feature”. We introduce a semantics divergence measure as:

$$ND_x^{a,b} = JS(\Theta_{a|ew}^x, \Theta_{b|ew}^x). \quad (8.4)$$

Named entities carry semantically rich information conveyed by the reports, but fail to capture the position of the reporters towards the story. News texts are often more than reporting, and express opinions and sentiments towards the story.

Sentiment divergence (SD). We therefore define a last measure based on the differences in used sentiment words. Since many words used to express sentiment are rarely used and the probability of observing them in a corpus is low, we follow the approach described in [1] and bin *sw* tokens into 7 categories of strength and type of the sentiment they express (ranging from strong negative to strong positive). Therefore, *SD* measures differences in probability distributions over the categories of sentiment, and not sentiment-bearing words:

$$SD_x^{a,b} = JS(\Theta_{a|bin(sw)}^x, \Theta_{b|bin(sw)}^x). \quad (8.5)$$

To be able to relate more than two sources to one another and to abstract from the (non-interpretable) absolute values of *JS*, we apply multidimensional scaling, projecting the obtained distance matrices into two dimensions.

8.5 Case study

We present a case study comparing news reports from Twitter (*tw*), blogosphere (*bl*), professional news outlets (*nw*), Reuters (*rt*) and Associated Press (*ap*).

Corpora and procedure. We obtained the story reports using the same query across different sources’ search engines. For news and blogs, we used Google News and Blog search to harvest web pages, and extracted content as described in [13]. We collected 3 breaking stories covering the BP oil spill, the Pakistan floods in 2010, and the Chilean miners’ accident, and 3 non-breaking stories about Belgian politics, Iraq, and the European Union. As an indicator of breaking stories we used Twitter’s trending topic list. The upper bound of corpus size was the number of tweets for the respective story, and the lower bound was set to 50. For collected documents we extracted named entities with Open Calais (www.opencalais.com), removed stop words, and lemmatized the rest.

Figure 8.1: Average RD for all stories comparing divergences between **Twitter** and other sources with *all* and *re,ap* baselines.

Figure 8.2: Average RD for **breaking stories** with the *non-breaking stories* baseline.

Relative divergence aggregation. The absolute values of KL based measures have no clear interpretation; we therefore concentrated on values relative to a baseline. We defined 3 baseline divergences. The first one (*all*) averages over all pair-wise divergences across all sources. The second base divergence value (*re,ap*) is the divergence between Reuters and Associated Press corpora. Due to the same type of media, format, and reporting style, we consider this as a reasonable baseline. To compare breaking and non-breaking stories, we used the average divergence for non-breaking stories over all sources as a baseline. We denote these 3 value as $baseline_{(all;re,ap;breaking)}$. The relative divergence (RD) of a source a for a story x is then:

$$RD_a^x = ((avg_{b \in sources} D_{a,b}^x - baseline_x) / baseline_x) \times 100. \quad (8.6)$$

Results. Figure 8.1. shows the results of applying RD to Twitter. We start with the interpretation of the *re,ap* baseline. The largest relative divergence is for sentiment divergence. The RD value of 67.87 shows higher sentiment divergence between Twitter and other sources. This result can lead to two conclusions: (a) Twitter contains more contrasting sentiment than news-wire reports, and (b) Twitter expresses more sentiment than news-wire reports. To decide between these two, we calculated the share of sentiment words across sources. In *ap* corpora, there are 1.7% sentiment words, in *re* corpora 2.8%, and in *tw* corpora 4.2%. We find that both the share and the type of sentiment words influence the differences between corpora. For example, strongly negative words make up 0.17% of the *ap* and 0.9% of the *tw* corpora. For other categories, it is expected that language divergence (LD) has a positive value, because more authors in Twitter use different writing styles, wording,

and non-standard grammar. This is partly shown by the average number of unique language-words tokens (10221 in *ap*, 13122 in *re*, and 89619 in *tw*). The high positive value of *HD* can be explained by the differences in the sizes of the different corpora, where a small number of documents in news-wire agencies do not converge to the same headline words. On average for a story, we collected 69 documents from news-wire agencies and 3314 from Twitter. The lowest *RD* value (-6.13) is found for named-entity divergences. This points to the conclusion that all sources are reporting on the same entities, but using different language and sentiment.

The difference between average *RD* values for the *all* baseline is always lower than for the *re, ap* baseline, on average $\approx 19\%$ lower. The extreme case is the difference of headline divergence values, which is 12.9 times lower when using the *all* baseline. We see this as an effect of having more documents to compare tweets to, because the average number of documents across all sources is 451, and due to many similar titles the divergence measure converges across sources. The lower *RD* values for *all* compared to *re, ap* suggest that Twitter is more similar to other sources than to news agencies.

Figure 8.2 shows *RD* values that describe the difference between breaking and non-breaking news. As a baseline divergence, we used the *non-breaking* value, comparing the average of the *breaking* stories to it. Positive values of *RD* reflect a higher divergence of reporting for breaking news. The figure 9.2 shows that breaking news is consistently more different across sources, except for sentiment divergence (*SD*). This suggests that for breaking news, informing the readers about the story is the main objective of the authors, while for non-breaking stories authors express their standpoints and analysis of the story.

To further investigate these differences and see which divergence contributes most, consider the MDS plots in Figure 8.3. Figure 8.3(a) shows that the headline distance between reports in News and Twitter is the lowest. Many news-related tweets come from Twitter accounts operated by professional news outlets [3]. In our dataset, we found an average of 2.9% identical entries in the *tw* and *nw* corpora. Figure 8.3(b) shows that Twitter uses language closer to news and blogs than news-wire agencies, while news and blogs use similar language when compared to other sources. In terms of named entities (Figure 8.3(c)), *re* corpora are far from other sources. This probably arises from the much number of named entities used by Reuters: an average per-document of 21.9 (compared to 0.22 in *tw*, 8.6 in *bl*, 12.91 in *nw*, and 13.2 in *ap*). Figure 8.3(d) visualizes sentiment divergence, showing that *bl, re*, and *nw* corpora contain similar amounts of sentiment, which are more different when compared to *tw* and *ap* corpora.

8.6 Conclusions and outlook

This work is our starting effort in defining an easily interpretable, multi-aspect similarity measures for comparing news sources. Of course, this work cannot cover all the possible or interesting aspects of divergence in news reports, and absolute values of divergence measures are hard to interpret. Nonetheless, as the paper has shown, the inspection of relative differences can give interesting insights, opening many interesting research directions.

We started this investigation by focusing on two roles of social media platforms: to create new and different news, or to peddle or spread existing news. In contrast to both, our results suggest that the biggest role of citizen journalists in news is the role of a *commentator*, not only reporting but expressing opinions and taking positions on the news. Investigating this role will yield further measures of relations between corpora and a deeper understanding of the dynamics of social media in today's news environment.

Figure 8.3: MDS maps of divergence measures: (a) *HD*, (b) *LD*, (c) *ND*, (d) *SD* .

References

- [1] Andrea Esuli Stefano Baccianella and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).
- [2] Alberto Barrón-Cedeño, Andreas Eiselt, and Paolo Rosso. Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions. In Dipti Misra Sharma, Vasudeva Verma, and Rajeev Sangal, editors, *In proceedings of the workshop on Comparing Corpora, held in conjunction ACL 2000. October 2000, Hong Kong*, pages 29–38, Hyderabad, India, 2009. Macmillan Publishers.
- [3] Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and P. Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.
- [4] Kenneth W. Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.
- [5] Peter E Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. Entropy Of English – An estimate of an upper bound for the. *Computational Linguistics*, 18:31–40, 1992.
- [6] Malcolm Gladwell. Small change: Why the revolution will not be tweeted. *The New Yorker*, 86(31), October 2010. web only, http://www.newyorker.com/reporting/2010/10/04/101004fa_fact_gladwell.
- [7] Lev Grossman. Iran protests: Twitter, the medium of the movement. *Time Magazine*, jun 2009.

- [8] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *Social Science Research Network Working Paper Series*, Dec 2008.
- [9] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.
- [10] Adam Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics*, 6:1–37, 2001.
- [11] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [12] John Lafferty and Chengxiang Zhai. Probabilistic relevance models based on document and query generation. In *Language Modeling and Information Retrieval*, pages 1–10. Kluwer Academic Publishers, 2002.
- [13] Jyotika Prasad and Andreas Paepcke. Coreex: content extraction from online news articles. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1391–1392, New York, NY, USA, 2008. ACM.
- [14] Clay Shirky. How social media can make history, Jun 2009. http://www.ted.com/talks/lang/eng-/clay_shirky_how_cellphones_twitter_facebook_can_make_history.html, TED talk.
- [15] Karin Verspoor, K. Bretonnel Cohen, and Lawrence Hunter. The textual characteristics of traditional and open access scientific journals are similar. *BMC bioinformatics*, 10(1):183+, 2009.
- [16] Dejin Zhao and Mary B. Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *GROUP '09: Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252, New York, NY, USA, 2009. ACM.

Errata

- Section 8.2, page 212, paragraph 2: *By some [14], this new way* should be Shirky [14] hails this new way;
- Section 8.2, page 212, paragraph 3: *social media provides a different reporting to* should be social media provides different reporting to;
- Section 8.2, page 213, paragraph 2: *complement the framework (Section 8.4) by a case study (8.5)* should be complement the framework (Section 8.4) with a case study (8.5);
- Section 8.3, page 213, paragraph 4: *[2] provides* should be Barron-Cedeno et al. [2] provide;
- Section 8.5, page 215, paragraph 5: *Twitter(*tw*), blogosphere (*bl*)* should be Twitter (*tw*), the blogosphere (*bl*);
- Section 8.5, page 216, paragraph 1: *We denote these 3 value as* should be We denote these 3 values as;