Peddling or Creating? Investigating the Role of Twitter in News Reporting

Ilija Subašić and Bettina Berendt

Department of Computer Science, K.U. Leuven, Belgium

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.

1 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 [16], 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 [9,7] 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 [8] and into the dynamics of content between news and blogs, e.g. [14]. We take a different approach: we start with news story already "discovered" and investigate whether social media provides a *different* reporting to traditional

P. Clough et al. (Eds.): ECIR 2011, LNCS 6611, pp. 207–213, 2011.

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 3) by a case study (4).

2 Related Work

Twitter research. Out of many areas of Twitter research, we focus on the ones related to news mining. [10,18,9] 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 [3,2] 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 longstanding topic in different fields producing a wide range of text similarity measures. [1] provides a valuable overview of different text similarity measures. Work in corpus linguistics [11] 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 [12], 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 reterieval [13], for detecting plagiarism in Wikipedia articles [1], and for comparing traditional and Open Access medical journals content [17]. We adopt a KL-based approach to corpora similarity, but provide multiple perspectives on similarity by combining several aspects of the corpora.

3 Measures of Corpora Divergence

Notation. A corpus C^{story}_{source} 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^{story}_{source} is a language model Θ^{story}_{source}, where the probability of a token t is denoted as $\Theta^{story}_{source}(t)$. 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 semantic entities (names, locations, companies, ...); and (4) sentiment words (sw): words expressing sentiment. $\varTheta^{story}_{source|category}$ denotes the category language model (e.g. for a unigram model, $\varTheta^{story}_{source|hw}$ 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 $\varTheta^x_a$ and $\varTheta^x_b$ to measure their distance as: $JS(\varTheta^x_a, \varTheta^x_b) = \frac{1}{2}KL(\varTheta^x_a \varTheta^x_m) + \frac{1}{2}KL(\varTheta^x_b, \varTheta^x_m)$ where the probability of every t in $\varTheta^x_m$ is the average probability of t in $\varTheta^x_a$ and $\varTheta^x_b$. 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: $LD^{a,b}_x = JS(\varTheta^x_{a|pw}, \varTheta^x_{b|pw})$. 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: $HD^{a,b}_x = JS(\varTheta^x_{a|hw}, \varTheta^x_{b|hw})$. 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: $ND^{a,b}_x = JS(\varTheta^x_{a|ew}, \varTheta^x_{b|ew})$. 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 [6] and bin sw tokens into 7 categories of strength and type of the sentiment they express (ranging form strong negative to strong positive). Therefore, SD measures differences in probability distributions over the categories of sentiment, and not sentiment-bearing words: $SD^{a,b}_x = JS(\varTheta^x_{a|bin(sw)}, \varTheta^x_{b|bin(sw)})$.

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.

4 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).

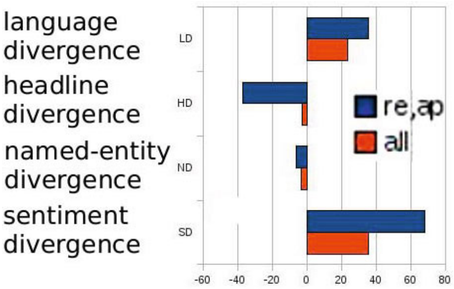

Fig. 1. Average RD for all stories comparing divergences between **Twitter** and other sources with *all* and re, ap baselines

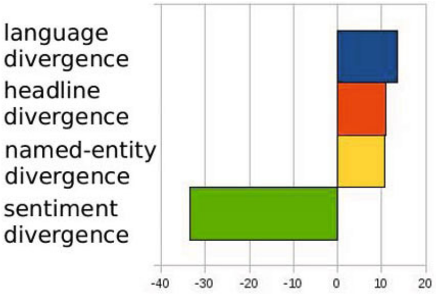

Fig. 2. Average RD for **breaking stories** with the *non-breaking stories* baseline

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 [15]. 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.

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: $RD^x_a = (\ (avg_{b \in sources} D^x_{a,b} - baseline_x)\ /\ baseline_x\ ) \times 100$.

Results. Figure 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 betweeen 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

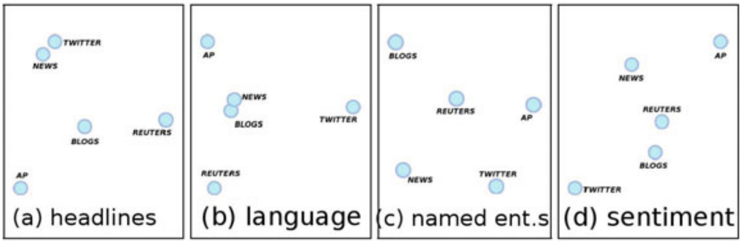

Fig. 3. MDS maps of divergence measures: (a) HD, (b) LD, (c) ND, (d) SD

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 ≈ 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 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 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 Fig. 3. Figure 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 [2]. In our dataset, we found an average of 2.9% identical entries in the *tw* and *nw* corpora. Figure 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 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 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.

5 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.

References

1. Barrón-Cedeño, A., Eiselt, A., Rosso, P.: Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions. In: Proc. of ICON 2009, pp. 29–38. Macmillan, Basingstoke (2009)
2. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring User Influence in Twitter: The Million Follower Fallacy. In: Proc. of ICSWM 2009. AAAI, Menlo Park (2009)
3. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in twitter: The million follower fallacy. In: Proc. of ICWSM 2010. AAAI, Menlo Park (2010)
4. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1), 22–29 (1990)
5. English, E.O., Brown, P.E., Pietra, V.J.D., Mercer, R.L., Pietra, S.A.D., Lai, J.C.: An estimate of an upper bound for the entropy of English. Computational Linguistics 18, 31–40 (1992)
6. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proc. of LREC 2006. LREC (2006)
7. Gladwell, M.: Small change: Why the revolution will not be tweeted (2010). New Yorker Magazine (October 2010)
8. Grossman, L.: Iran protests: Twitter, the medium of the movement. Time Magazine (June 2009), http://www.time.com/time/world/article/0,8599,1905125,00.html
9. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. ArXiv e-prints (December 2008), http://arxiv.org/abs/0812.1045

10. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proc. of WebKDD/SNA-KDD 2007, pp. 56–65. ACM, USA (2007)
11. Kilgarriff, A.: Comparing corpora. International Journal of Corpus Linguistics 6, 1–37 (2001)
12. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics 22(1), 79–86 (1951)
13. Lafferty, J., Zhai, C.: Probabilistic relevance models based on document and query generation. In: Language Modeling and Information Retrieval, pp. 1–10. Kluwer, Dordrecht (2002)
14. Leskovec, J., Backstron, L., Kleinberg, J.M.: Meme-tracking and the dynamics of the news cycle. In: Proc. of KDD 2009, pp. 497–506. ACM, New York (2009)
15. Prasad, J., Paepcke, A.: Coreex: content extraction from online news articles. In: Proc. of CIKM 2008, pp. 1391–1392. ACM, New York (2008)
16. Shirky, C.: How social media can make history. TED Talk (June 2009), `http://www.ted.com/talks/clay_shirky_how_cellphones_twitter_facebook_can_make_history.html`
17. Verspoor, K., Cohen, K.B., Hunter, L.: The textual characteristics of traditional and open access scientific journals are similar. BMC bioinformatics 10(1), 183+ (2009)
18. Zhao, D., Rosson, M.B.: How and why people twitter: the role that micro-blogging plays in informal communication at work. In: Proc. of GROUP 2009, pp. 243–252. ACM, New York (2009)