Linking Tweets to News: Is All News of Interest?

Tariq Ahmad(✉) and Allan Ramsay

School of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK
tariq.ahmad@postgrad.manchester.ac.uk, allan.ramsay@cs.man.ac.uk

Abstract. In a world where news is being generated almost continuously by many different news providers on many different platforms, it would be useful in certain industries to be able to determine how much of that news is actually being read, which news items are not interest generating or, indeed, if there are topics being discussed on Twitter that have not even been reported in the news. Twitter generates vast numbers of Tweets daily and has a massive active user base, so it is ideal as a way of gauging what news people are, or are not, interested in. This paper proposes a technique to efficiently relate Tweets to news articles and then to determine which news articles are of interest, which are not, and what is being discussed on Twitter that is not even in the news.

Keywords: Twitter · News articles · Similarity · TF-IDF

1 Introduction

Twitter generates 500 million Tweets daily [4] and has a 320 million active monthly user base [8]. Many of these Tweets will talk about a current news topic, many of them will refer to something that has not appeared in the mainstream news yet and many will be "noise". If we could link news articles to Tweets we could perhaps determine which news items are generating interest and which are not. Tweets are random in nature, in that they may contain unstructured text, abbreviations, slang, acronyms, emoticons, incorrect grammar and incorrect spelling [1]. When we took individual Tweets and tried to link them to news articles using simple word matching this produced inaccurate matches. This is because individual Tweets do not contain enough information to perform this task. When we tried to match 10,000 Tweets to 11,888 news articles, using simple word matching alone (excluding hashtags, URLs and usernames) we had 9 % returned as matching. However when we looked at the results, a Tweet such as *"@TekelTowers @mygransfortea not mock"*, is clearly not related to the suggested news article headlined *"Donald Trump says he did not mock New York Times reporter's disability"*. Finding Tweets that are actually related to news stories, rather than finding ones that share low frequency terms with them, is a challenging task.

The problem is that individual Tweets are very short, and hence TF-IDF vectors for them will be heavily swayed by specific terms. The aim of the work

C. Dichev and G. Agre (Eds.): AIMSA 2016, LNAI 9883, pp. 151–161, 2016.
DOI: 10.1007/978-3-319-44748-3_15

reported here is to group Tweets into larger collections, which we will refer to as "stories", in order to downplay the effects of individual terms. If the Tweet above about going to gran's for tea had been linked into the larger body of material about daily life which is posted by *@TekelTowers* then the presence of the word *'mock'* would have been masked by information about Scottish football, and this Tweet would not have been linked to material relating to Donald Trump.

Tweets include two devices for grouping them – usernames and hashtags. Usernames are helpful if there is a particular individual whose opinions you care about, but as [2], among others, has noted this does not tend to lead to thematically linked material. Hashtags, however, highlight the topics in Tweets [2], e.g. *"@Vistaprint are worse then @Ryanair with their shopping cart experience #tiring"* and hence two or more Tweets sharing the same hashtag should be related. We therefore use Twitter hashtags and usernames to group Tweets into "topics that people are talking about". We refer to these as "Tweet stories".

News articles tend not to have such strong indicators. They do, however, tend to contain large numbers of words, and hence TF-IDF vectors do provide a robust way of grouping them into stories, particularly if we also make use of the time-stamps associated with entries on RSS feeds.

The final step in the work reported here is to link Tweet stories to RSS stories to establish which news stories are being talked about on Twitter, and by elimination which news stories are **not** being talked about on Twitter and which Twitter stories are not showing up in the news. Tweet stories that are not linked or loosely linked to news stories are deemed as "things that people are talking about, that are not in the news". This exercise will yield a number of different outcomes:

1. Topics that appear in news articles and are being discussed on Twitter
2. Topics that appear in news articles but are not being discussed on Twitter
3. Topics that do not appear in news articles but are being discussed on Twitter

There are a number of different players such as news providers, governments, politicians or data analysts who could potentially benefit from knowing about the relationship between news stories and what is being discussed on Twitter.

News providers would be interested in knowing what topics are in the news that people are not talking about on Twitter because that might draw their attention to the fact that they may be spending substantial resources on producing content that, actually, is not of interest to anyone. They might also be interested in knowing that there are things that people are discussing on Twitter that they are not reporting on. For example, if there was a lot of discussion on Twitter about the weakness of Putin but this was not reported in the news, this might be of considerable interest to certain types of people. In general, as discussed by Howard et al. [3], countries with repressive state-controlled media would certainly be interested to know that there are things being discussed on Twitter that might precede major events on the ground.

Twitter also moves faster than platforms for traditional news dissemination. If we could find patterns in Tweet stories that indicated that they were

likely to turn into news stories, we could act more quickly in the face of, for instance, epidemics, which might show up on Twitter as informal observations about symptoms and absence from work well before there was reliable epidemiological evidence.

2 Data Collection

We need a large number of Tweets. Given that we need to use time-stamps as part of the process of linking individual Tweets and articles into Tweet stories and news stories, and that we then need to use this same information to constrain the links between Tweet and news stories, it is easier to construct a corpus with the required characteristics than trying to find existing corpora that cover the same time span and hence can reasonably be expected to contain linked stories.

We therefore gathered Tweets originating in the UK, with no restriction on topic, and news articles from RSS feeds available from a number of major news providers with UK outlets, namely Google, BBC, The Independent, The Guardian, The Daily Express, Sky and CNN, over a period of 30 days. The final dataset contained 693,527 Tweets and 11,896 news article webpages.

We did not restrict the news articles or Tweets to any specific topic, keyword or news event. The aim of the project is to find out what kinds of Twitter stories are under- or over-represented in the traditional news media: making any kind of prejudgement about what area we want to look at would inevitably compromise this goal. We may have some initial thoughts about the likely outcome of this investigation, but it would be a mistake to build those thoughts into the data-gathering process.

3 Stories

Documents that are published in the same time interval have a larger chance of being similar that those that are not chronologically close [2]. However, we cannot simply assume that documents are similar based on their timestamp.

3.1 RSS Stories

We use cosine similarity to group our news articles. Guo et al. [2] represent a news article using its *summary* and its title, however we represent a news article by using its *content* and its title.

The advantage of using the summary is that it is an easily identifiable piece of text, without extraneous links to other articles, adverts and external websites, and without a large amount of embedded HTML. The disadvantage is that it is, indeed, a summary, and hence contains less information than the full article.

Extracting the content from the content of an article poses a number of challenges, because RSS feeds tend not to follow a consistent format, and the content of the article is often buried inside a large amount of irrelevant material.

However, a reasonable rule of thumb is that consecutive passages of material separated by plain <p> tend to be content, whereas things like adverts, links to other pages, and side-bars tend be contained inside other kinds of tags. Not all news providers follow this pattern, but for the ones noted above this is a reliable way of finding the content.

We construct stories as follows. Initially, we have no stories. We examine the first document D1 and perform cosine similarity against all the other documents that are grouped into stories that were last updated in the last two days, and look for the story that has the highest similarity score above a similarity threshold. Since D1 is the first document we have ever seen, we do not have any matching stories. Therefore this document forms a story of its own containing just the one document D1. When we add a document to a story we update the story TF matrix that we will use to compare subsequent documents. We now look at the next document D2 and again compare it to all documents that are grouped into stories, using the story TF matrix. If D2 is similar enough to S1, then D2 is added to story S1, otherwise it forms a new story S2. This process is repeated until all documents for Day 1 have been classified into stories or they are left unclassified because the similarity measure did not exceed the threshold value for any story.

For Day 2 we follow the same procedure, but with one crucial difference. As well as comparing documents from Day 2 with stories created on Day 2, we also compare documents from Day 2 with stories created on the previous day (i.e. Day 1), this gives us the concept of "running stories", i.e. stories about the same topic that appeared in the news over multiple continuous days.

The effectiveness of this strategy depends on choosing a suitable threshold for deciding whether an article should be included in a running story or should form the basis of a new story by itself. If the threshold is too high, every article forms a story by itself, if it is too low then articles that are not linked get grouped into the same story. Given that we are interested in assessing the importance of the news stories that are and are not linked to Twitter stories, it is important to get this right, since the most obvious measure of the importance of a story is the amount of space that is devoted to reporting it, and we can only measure that accurately if we have indeed grouped articles appropriately.

To determine a trustworthy, reliable threshold we conducted an online survey. We selected 100 random news article pairs with cosine similarities in the range 0.5–1.0, ensuring that we had 20 pairs with similarity 0.5–0.6, 20 pairs with similarity 0.6–0.7 and so on. Furthermore, we ensured that the survey presented pairs that had been answered the least first to ensure that we had answers for as many of the pairs as possible. The survey presented four users with the headline and short description from two related articles and asked them to say if the two were related or unrelated. The results are presented in Table 1.

The results show that when the cosine similarity was from 0.9–1.0, 92 % of the time the users agreed with our classification that the articles were related. However, as discussed, too high a threshold is problematic. When the cosine similarity was from 0.8–0.9, 70 % of the time the users agreed with our classification.

Table 1. Results from survey

Pair cosine similarity	Total questions answered	Unrelated (%)	Related (%)
0.9–1.0	79	7.59	92.41
0.8–0.9	78	29.49	70.51
0.7–0.8	78	55.13	44.87
0.6–0.7	78	53.85	46.15
0.5–0.6	78	66.67	33.33

Below 0.8 we see that our classification was correct less than 50 % of the time, so it is not sensible to use a threshold less than this as this would lead to unrelated articles being grouped into stories.

0.8 would seem to be the best value to use. However, to ensure that this inter-annotator agreement was not simply due to chance we used Fleiss' kappa to get a kappa value of 0.69, which according to the commonly cited scale as described in [6] means that there is "Substantial agreement" between our annotators.

We therefore decided that 0.8 was a reliable value to use as our threshold.

3.2 Tweet Stories

We use hashtags to group our Tweets. The concept is similar to that employed for RSS stories, in that we create stories and then turn them into running stories as appropriate. Even though we have vastly larger amounts of data than for our RSS stories, this process is much more efficient and can process large numbers of Tweets easily. One slight difference in the creation of Tweet stories, is that we do not allow Tweet stories to contain only one Tweet. This is because, as we have seen, it is very easy for an individual Tweet to get falsely linked to a story on the basis of a few keywords, but it is much harder for a Tweet story to falsely get related to a news story. That is not to say it will **never** happen, just that it is much harder for it to happen. As in [10], any Tweets not grouped with other Tweets are discarded. We find that these are typically short, single, random Tweets such as *"Just spilt milk on my laptop!!! Help!!!"* or *"@StephenChamber8 no I love them"*. This is not the content "that does not appear in news articles but is being discussed on Twitter", these are single, random Tweets.

4 Linking Tweet Stories and RSS Stories

We have now managed to group together Tweets into Tweet stories and RSS news articles into RSS stories. The numbers of Tweets has been reduced by approximately 33 % but the number of RSS documents remains unchanged because we allow RSS news stories to contain one or more items.

We now want to link the Tweet stories to the RSS stories. We would like to do this using cos_{TF-IDF}. However some stories run for days which leads to

Table 2. Numbers of documents before and after grouping into stories

	Initial	After grouping	Stories
RSS	11,896	11,896	11,457
Tweet	693,527	462,212	102,938

the construction of TF matrices which contain thousands of words. This gives us a problem in that we find that as the number of words increases, the cosine similarity algorithm gets slower (Table 3).

Table 3. Cosine similarity computation comparison

Entity	Average words	Time (s)
Tweet	12	0.00003
Document	392	0.0002
Story	3371	0.001

In order to find the best match for each of our Tweet stories in the set of RSS stories we would need to perform 102,938 × 11,457 cosine similarity computations. This takes of the order of tens of hours to compute. As others have discovered [7], this is impractical, even if we restrict attention to stories that have overlapping times (remember that Tweet stories often start before the corresponding news stories, so we cannot simply focus on pairs that cover the same time periods).

4.1 Top 5 Words

We know that there will be lots of stories that will be unrelated to each other. We would like to exclude these from the computation-intensive cosine similarity part of our process because that would be wasting time and effort. Ideally, we need a quick way of finding stories that **might** be linked, and then we can use cosine similarity to confirm or refute that.

The main idea in this paper is that we create cosine similarity "candidates" by determining the top five words from each story using TF-IDF and then simply linking stories that share one or more top words. If two or more stories mention the same important keywords then we hypothesise that they might be describing the same event, and then we use cosine similarity on these to determine whether these stories are really related or not.

One might think that because stories share keywords that they would almost certainly be linked, but this is not the case. A typical example of this scenario is where the word "quiz" is flagged as a top five word in a Twitter story and in an RSS story, but upon closer inspection we see that the Tweets are to do

with *"Jimmy Carrs Big Fat Quiz of the Year"* whereas the RSS story is about a suspect who was "quizzed" by the police. Clearly these two stories should not be linked, cosine similarity will ensure that is the case. This process gave us 2,536 candidates between Twitter and RSS stories. We then performed cosine similarity on these to get a measure of how similar the pair is.

5 Results

5.1 Tweet Stories Not Related to News Stories

One of the immediately obvious results is that the number of Tweet stories has been reduced from 102,938 to 2,536 stories that relate to an RSS news story.

From the original 102,938 stories we find that 72,191 are stories that were created using usernames alone (e.g. *@LevityMusic*). These stories contain a varying number of Tweets, from 2–259, with only 3,688 stories containing ten or more Tweets. These are discarded as discussed.

The remaining 28,211 stories are those for which we are unable to find related RSS stories - these are the topics that are being discussed on Twitter but do not appear in the news (Table 4).

Table 4. Tweet stories results summary

Result	%
Based on username	70.1
Not related to a news article	27.4
Related to a news article	2.5

The 28,211 stories that are not related to news articles also have varying numbers of Tweets ranging from 2–2,319, with only 2,212 stories containing ten or more Tweets. The story with 2,319 Tweets was about *#Sherlock*. Other unmatched stories containing large numbers of Tweets were about *#HappyNewYear*, *#EthanAndGraysonTo1Mil*, *#LoveTheDarts*, *#BoxingDay* and *#2016*.

Some of the Tweet stories that contained 100 s of Tweets were about things like *#corrie*, *#SutthakornTour*, *#Lotto5Millionaires* and *#2015Wipe*.

At the bottom end of the scale where we have stories containing less than ten Tweets we see hashtags like *#10SpinHitsRewind*, *#hungryeyes*, *#band*, *#mmromance* and *#independentwomanbeautyclinic*.

The stories with high numbers of Tweets seem to be about topics (Darts, Sherlock, New Year) that we would think, and probably are, being reported on in news articles. However, the *#EthanAndGraysonTo1Mil* hashtag looks like exactly the kind of thing we are looking for - it has a high number of Tweets but a Google search on *"Ethan And Grayson To 1 Million"* reveals only one relevant news article.

5.2 News Stories Related to Tweet Stories

From the numbers of stories listed in Table 2, we only manage to produce 2,536 supposedly related stories. However, the cosine similarities of these are 0.24 or below. We expected to get more closely matched stories due to the steps we performed above. We know that Tweets are short, in comparison to RSS news articles, therefore Tweet stories are also short in comparison to RSS news stories. In our dataset, on average, a Tweet contains 12 words whereas a news article contains 770 words.

Extrapolating this we can see that this mismatch in numbers of words increases for stories. We know cosine similarity uses the concept of treating words as vectors, if vectors are close to each other then they are similar. The reason our similarity measures are small is because our RSS stories have a high number of dimensions compared to our Tweet stories and, although the RSS vector will be far away from the corresponding Tweet vector, the keywords we have identified will help to ensure that we get at least some sort of positive similarity measure.

We expect that the items we found will be closely related, however on first inspection this does not appear to be the case. The supposed best match we found (with a similarity score of 0.24) was a false-positive between a news article titled *"Obama confident executive action on gun control is constitutional"* and Tweets like *"Liberty London ??? #libertylondon #beautiful #london @Liberty London"*. These are clearly not related. However, the next best score of 0.21 is between an article titled *"No reason to resign - Van Gaal"* and Tweets like *"Lous van Gaal sat motionless, what an absolute shock! #mufc #GiggsIn #LVGOut"*. These, clearly, are related.

5.3 News Stories Not Related to Tweet Stories

Recall, that originally we had 11,896 RSS news articles. These were then grouped into 11,457 stories and then further processed into 2,536 RSS-Tweet candidate pairs. From this process we find that there are 158 RSS stories that did not get matched to a Tweet story as part of the "create candidate pairs" process. Out of these there are 43 that contain more than two documents. If we look at the content of some these we see headlines such as *"Paris attacks: 5000 rounds fired in St-Denis raids, prosecutor says"*, *"Global climate march 2015: hundreds of thousands march around the world"*, *"Jihadi John'dead': MI5 on alert amid fears of Isil revenge attack"* and *"Failed flood defences cast doubt on UK readiness for new weather era"*.

These are definitely stories that we would reasonably expect to see discussed on Twitter. We can perhaps surmise that we failed to match these because of our mismatch in collecting worldwide articles against UK Tweets.

Alternatively, as mentioned by Zhao et al. [9], this could be because Twitter users show relatively low interest in world news.

Gold Standard. We selected 400 spread out pairs from the upper end of our results and hand marked them as "Correctly identified" or "Incorrectly identified" to get Precision and Recall graphs.

The Precision graph shows that the pair with the highest score was actually a mismatch, then between a score of 0.10 and 0.22 we see that a high proportion of the matches we are seeing are correct. Between 0.15 and 0.10 we see the matches we find are correct only 50 % of the time. Beyond a score of 0.10, as we might expect with such low scores, there are fewer and fewer correct matches.

Our Recall graph shows that beyond a certain small score there are no further matches to be found.

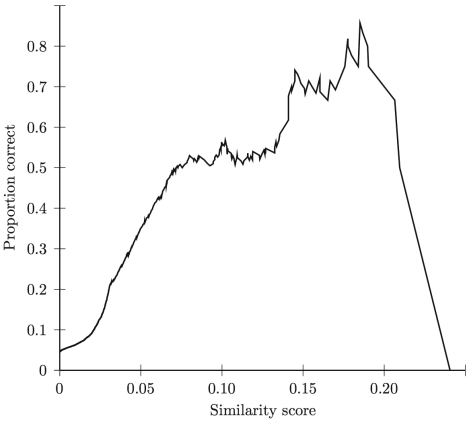

Fig. 1. Precision

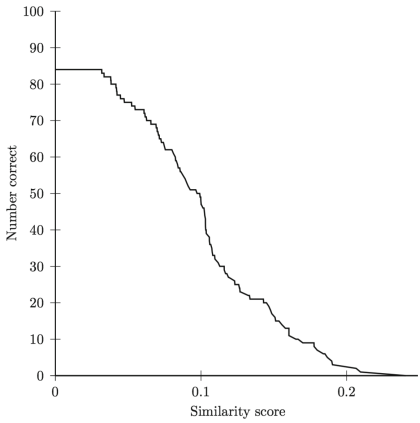

Fig. 2. Recall

6 Related Work

There are many techniques that have been proposed for grouping "documents" (i.e. news stories, Tweets or indeed any kind of body of text).

Kaur and Gelowitz [5] proposed using cosine similarity to group Tweets. Although this approach grouped Tweets effectively, they only used a relatively small number of Tweets (3,500). For the numbers we used, cosine similarity quickly became impractical as we saw. Furthermore, they remove punctuation, URLs, usernames and also make use of a stoplist. We prefer to let the mathematics of TF-IDF take care of common words like "the" and "and" that do not tell us anything.

In [10], Petrović et al. also link Tweets to news articles using cosine similarity but they do not use a threshold as we do.

Similar to our work, Sankaranarayanan et al. [11] also use cosine similarity in conjunction with publication time - but they use the mean publication time of Tweets in the story.

The work done by Guo et al. [2] is also a similar concept in that they attempt to link Tweets to news, but they try to predict the URL referred news article based on the text in each Tweet.

We also find, as discussed by Zhao et al. [9], that Twitter is a good source of topics that have low coverage in the traditional news media.

7 Conclusion

It has been shown that by creating cosine similarity candidates by using the top five words in each story, we can filter out many of the stories that were unlikely to match. This then allows us to perform cosine similarity in an efficient manner on the remaining pairs.

At a later stage, in another work, we plan to find informal sentiment expressions in Tweets. We plan to do that by looking for orthodox expressions or sentiment markers in news sources and looking for over-representation of particular terms in Tweet stories. This project also serves as part of our motivation for that project.

Acknowledgments. Tariq would like to thank the Qatar National Research Foundation for their financial support. Allan Ramsay's contribution to this work was also supported by the Qatar National Research Foundation.

References

1. Albogamy, F., Ramsay, A.: POS tagging for Arabic tweets. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, pp. 1–8. INCOMA Ltd., Shoumen (2015). http://www.aclweb.org/anthology/R15-1001
2. Guo, W., Li, H., Ji, H., Diab, M.T.: Linking tweets to news: a framework to enrich short text data in social media. In: ACL (1), pp. 239–249. Citeseer (2013)
3. Howard, P.N., Duffy, A., Freelon, D., Hussain, M.M., Mari, W., Mazaid, M.: Opening closed regimes: what was the role of social media during the Arab spring? In; Social Science Research Network (2011)
4. internetlivestats: Twitter usage statistics. http://www.internetlivestats.com/twitter-statistics. Accessed Mar 2016
5. Kaur, N., Gelowitz, C.M.: A tweet grouping methodology utilizing inter and intra cosine similarity. In: 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 756–759. IEEE (2015)
6. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
7. Petrović, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to Twitter. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT 2010, pp. 181–189. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
8. Twitter: About Twitter. https://about.twitter.com/. Accessed Mar 2016

9. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X.: Comparing Twitter and traditional media using topic models. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 338–349. Springer, Heidelberg (2011)
10. Petrović, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., Shrimpton, L.: Can Twitter replace newswire for breaking news? In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (2013)
11. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: news in tweets. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 42–51 (2009)