Topic and Emotion Evolution in News: Characterisations and Measures [*]

Quang Minh Nguyen
School of Electrical Engineering

Maida Aizaz
School of Computing

1 Introduction

News data is abundant and contains rich information about the world. As such, analysing it in terms of topics and emotions can provide significant insights. In fact, such analysis has been carried in a large body of literature in different contexts, yielding many meaningful applications. For example, topic coverage and dynamics can be utilised as a tool for investigating emerging issues: [6] explored how the Ebola outbreak in 2016 was manifested in the evolution of news on Twitter and in the news. The analysis of news sentiment can be applied to administrative decision making [7]. Another application studied was the degree to which future popularity or market behavior are influenced by sentiment in news and blogs [5].

In order to successfully capture the rich information contained in news data, it is important to consider temporal dynamics, as suggested by the aforementioned past work. This motivates the research question that we pose: *what is the relation between topic evolution and emotion evolution*? To answer this question, we define novel characterisations for the evolution of topics and emotions, as well as a measure for their difference, which are described below. We make the code and datasets used in this work publicly available at https://github.com/ngqm/topic-emotion-coevolution. Following our detailed analysis, we come to the conclusion that *topics and emotions in news have different evolution patterns.*

In the next section, a brief literature review is provided. In section 3, we formally describe important concepts and give information on our dataset. We characterise the evolution of topics and emotions and compare them with our measure in section 4. In section 5, we interpret our findings. We conclude our research and propose directions for further work in section 6.

2 Background

There has been significant work in the past on topic and sentiment analysis, especially in the domain of news, using a variety of different strategies.

- One approach is to consider topic and sentiment analyses independently from each other. In particular, Fukuhara et al. conducted temporal sentiment and topic analysis on articles from a Japanese newspaper [4]. Balahur et al. performed a similar sentiment analysis on quotes extracted from NewsBrief and MedISys [1].
- Another approach discusses sentiments in topic-based contexts. O'Hare et al. were amongst the first to do this, conducting their analysis on financial blogs [10].

Our focus is on temporal topic and sentiment analysis, particularly so in news. We call this topic and sentiment *evolution.* Liu et al. proposed a dynamic model for large datasets of online news publications as well as readers' comments on them, their goal being extraction and analysis of topic-based sentiments, as well as their evolution over time [8]. In a similar manner, Chen et al. used Latent Dirichlet Allocation (LDA) and cosine similarity to investigate topic evolution

[*] Coursework in HSS312 Computational Social Science, offered in Fall 2021 at KAIST

in a scientific domain [3]. We add to the final approach by contibuting a new method for analysing topic and emotion evolution.

Some authors use the term *emotion* to refer to both sentiments (general attidudes, i.e., positive or negative) and emotions (particular feelings, e.g., fear or happiness), whilst others include both emotions and sentiments in the term *sentiments*. In this work, we use the former definition of *emotion*.

3 Method

In this section, we (1) formally describe the concepts on which our analysis is built on, (2) state the goal of our analysis, and (3) provide information on the utilised dataset.

3.1 Topic and emotion temporal similarity tensors

We first describe how a topic and its associated emotions are represented. Then comes our theoretical contribution, the temporal similarity tensors for topics and emotions.

Hereafter, when refering to a vocabulary, we mean a set of words. A generic vocabulary V is the set of all words in English, which is the notation we will use from now on. The emotion vocabulary E is the set of 10 emotions and sentiment, according to the NRC lexicon [9],

$$E = \{\text{anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative, positive}\}\,. \tag{3.1}$$

Definition 3.1 (Topic)**.** *A topic over a vocabulary V is a probability distribution over V. When the context is clear that V contains all English words, we omit the vocabulary and simply say **topic**. The probability of a word w in a topic t is denoted by $t(w)$.*

A lexicon associates each word in a vocabulary with a subset of the emotion vocabulary. This means a word might have none, one, or many associated emotions. In this work, we follow the associations made by the NRC lexicon. For a word w and an emotion e, $w(e)$ is 1 if w is associated with e and 0 otherwise.

Definition 3.2 (Topic-induced emotion)**.** *The emotion induced by a topic t is a probability distribution over the emotion vocabulary E, where the probability for an $e \in E$ is computed as*

$$ed(t, e) = \frac{\sum_{w \in V, w(e)=1} t(w)}{\sum_{e' \in E} \sum_{w \in V, w(e')=1} t(w)}\,. \tag{3.2}$$

Here *ed* is for *emotion distribution.*

In our analysis, 10 topics will be extracted per time period on a temporal text data set, using Latent Dirichlet Allocation. We denote this topic data by $t_i^{(k)}$, where $k = 1..N$ is the time index and $i = 1..10$ is the topic index. We also denote induced emotions in the same manner, replacing t with e, into $e_i^{(k)}$.

Note that while t and ed are probability distributions, we can encode them into unique vectors by fixing the order of words/emotions. In our research, the order of words is according to the dictionary order, and the order of emotions is according to 3.1. The temporal similarity tensors can then be defined as

Definition 3.3 (Temporal similarity tensors)**.** *Given the temporal topic data $\{t_j^{(i)}\}$, we define the topic similarity temporal tensor TS as*

$$TS_{kij} = \frac{t_i^{(k)} \cdot t_j^{(k+1)}}{\|t_i^{(k)}\| \|t_j^{(k+1)}\|}\,, \tag{3.3}$$

where $k = 1..n-1$ is the time index, and (i, j) is the index for an ordered pair of topics from months k and $k+1$. We also define the induced emotion similarity temporal tensor ES as

$$ES_{kij} = \frac{e_i^{(k)} \cdot e_j^{(k+1)}}{\|e_i^{(k)}\| \|e_j^{(k+1)}\|}\,, \tag{3.4}$$

where we employ the same index notation as above.

The temporal similarity tensors makes use of cosine similarity. This is a widely used similarity metric in the field of natural language processing. These tensors capture the *evolution* of topics and their emotions: information regarding the change of topics and emotions each month is recorded.

Although we make use of the temporal similarity tensor, in subsequent results and discussions, the name *topic similarity* and *emotion similarity* will also be used interchangeably for convenience. The former refers to entries in the topic tensor and the latter refers to entries in the emotion tensors.

3.2 Comparing topic and emotion evolution

Now that the evolution of topics and emotions have been represented as temporal similarity tensors, we can compare between the two evolution processes. This comparison is formalised using the $L1$ distance between corresponding slices of the two tensors, which we will call evolutional difference.

Definition 3.4 (Evolutional difference)**.** *Given topic and emotion temporal similarity tensors TS and ES, we define the evolutional difference at time k as*

$$D(k) = \frac{1}{100} \sum_{ij} |TS_{kij} - ES_{kij}| \,, \tag{3.5}$$

where $\|.\|$ denotes absolute value.

The absolute difference is divided by $100 = 10 \times 10$ here to account for 10 topics each period.

3.3 Dataset

We extract topics and emotions from the open dataset *All the news*, consisting of 146032 news articles from 15 American publications. The data is largely dated between January 2016 and June 2017, which is the period to be analysed. We only make use of the text content of each article, ignoring the titles. The period unit we consider is month. There are accordingly 18 such periods.

It is inevitable that a publication may be offensive, on purpose or accidentally, against specific groups. However, we do not filter out such materials because of two reasons: (1) the borderline between unethical texts and the ethical counterpart is blurry and highly subjective, and (2) the inclusion of all available articles ensures truthful observations of real-world news and allows us to draw accurate conclusions about the structure of our data.

The article count each month is included in Figure 1. The number of articles from each publication is included in Figure 2.

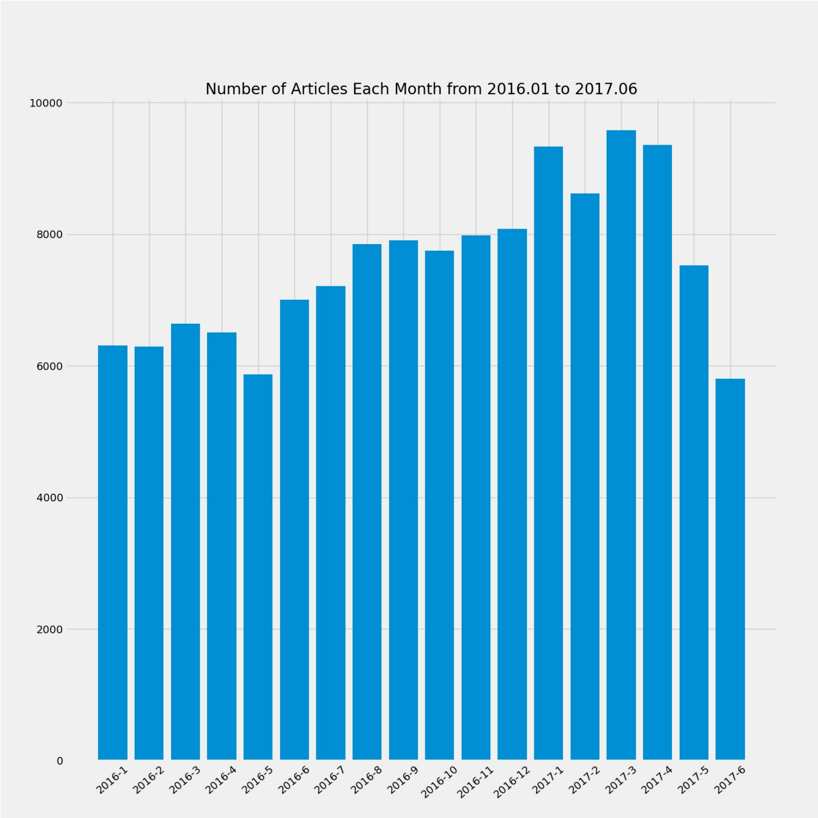

Figure 1: Number of articles each month from January 2016 to June 2017

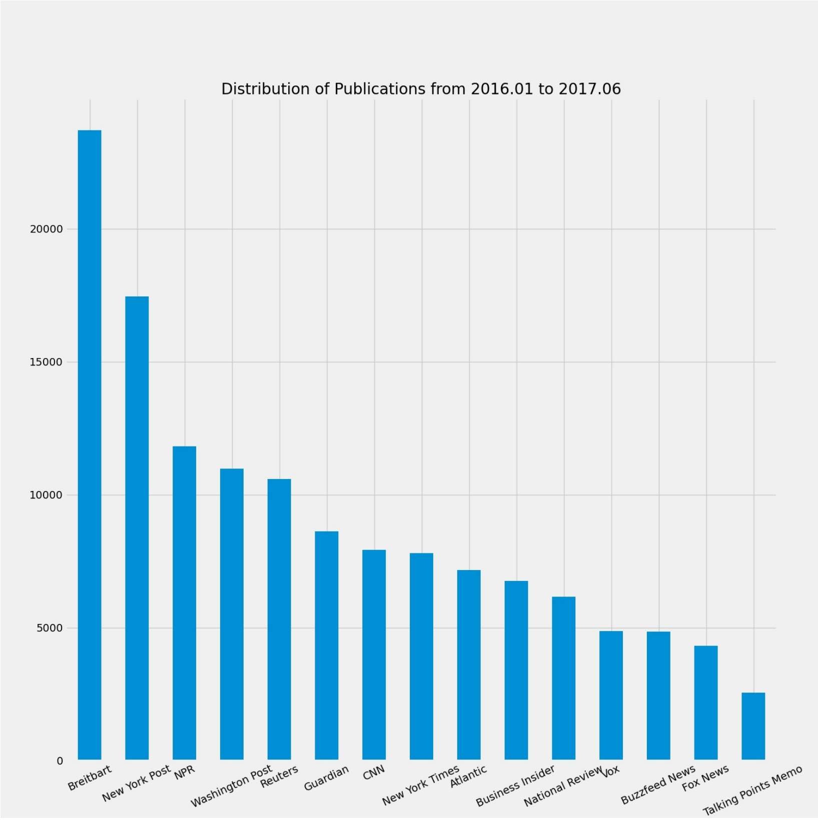

Figure 2: Number of articles from each publication from January 2016 to June 2017

4 Results

4.1 Topic extraction using Latent Dirichlet Allocation

The 10 extracted topics with their associated distributions can be visualised as shown in Figure 3, where we displayed the topics extracted from January 2017 as an example. Topic 2 here is evidently a popular topic, appearing often in the articles. However, our research does not take into account the popularity of various topics. Regardless of popularity, we normalise all topics to become a probability distribution, assigning each word a probability $t(w)$.

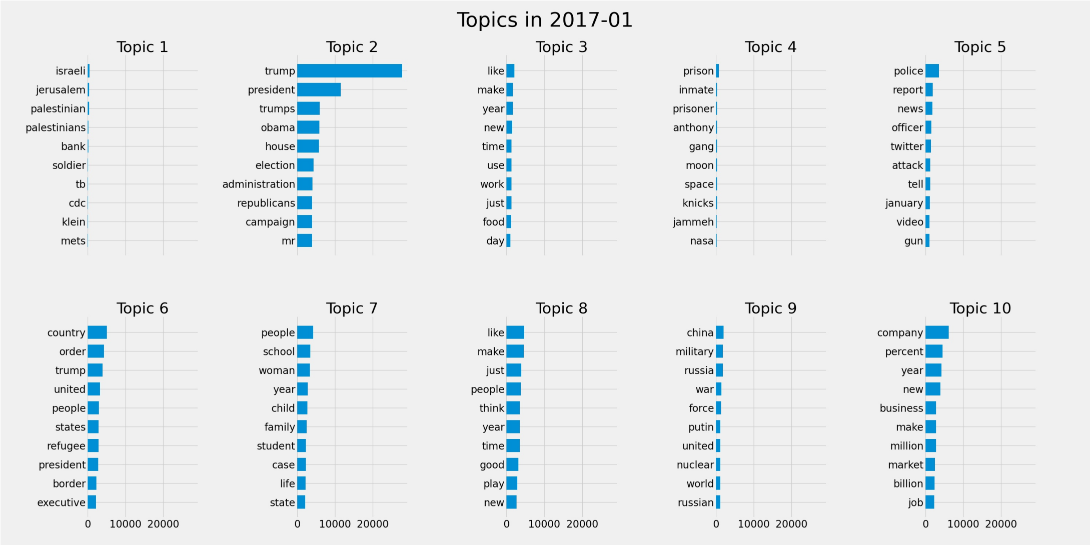

Figure 3: Topics extracted from January 2017

4.2 Topics and emotions evolve differently

In this part, we provide evidence that topics and emotions have different evolution patterns, using qualitative analysis and the evolutional difference metric.

First, the similarity between topics and emotions over the first 8 months is depicted in Figure 4.

- In each matrix of the first row, the entry at row i, column j represents the similarity TS_{kij}
- In each matrix of the second row, the entry at row i, column j represents the similarity ES_{kij}
- The darker the hue of particular element of the matrix, the higher the similarity observed.

It can be seen that the topics vary in similarity more than the emotions do; the hues of the topic matrices' elements show greater variation, although they tend to remain on the lower end of the scale. Meanwhile, most of the elements of the emotion matrices are dark, indicating very high similarity in emotions throughout the 8 months we consider. Furthermore, most topics are similar to only topic of the next month, as seen by the sparse distribution of singular dark elements.

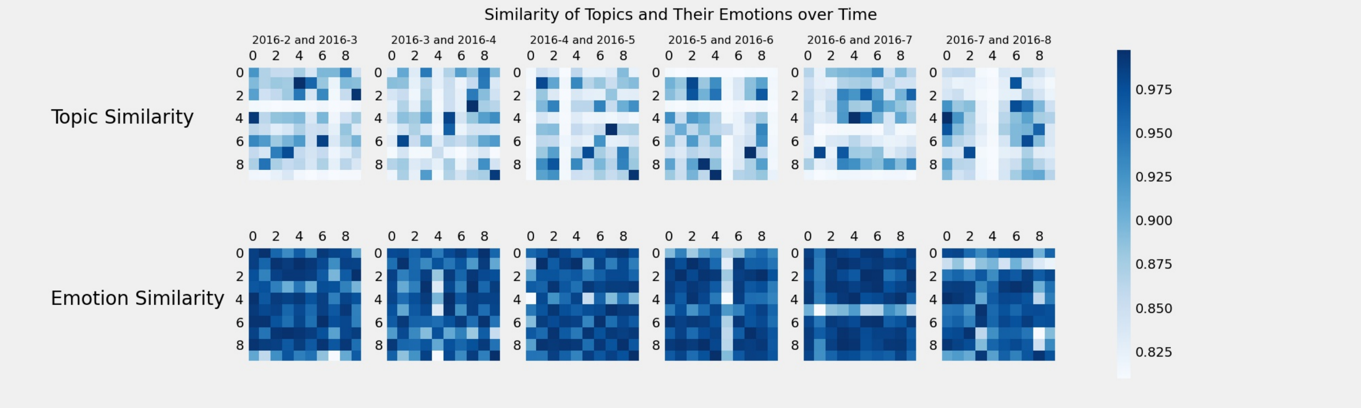

Figure 4: Similarity of topics and their emotions over time

Second, the correlation between topic similarity and emotion similarity is visualised as a series of scatterplots in Figure 5. Formally, for each k, we plot $(TS_{kij}, ES_{kij})_{ij}$. From this figure, we infer that for high topic similarity, the emotion is subsequently high but low topic similarity tends to have a wide range of emotion similarity–a pattern consistent across each pair of months. A combined scatterplot of all 18 months is shown in Figure 6, which corroborates the aforementioned inference. In this figure, it is further evident that emotion similarity is distributed at a very high range, different from topic similarity–a remark we made earlier.

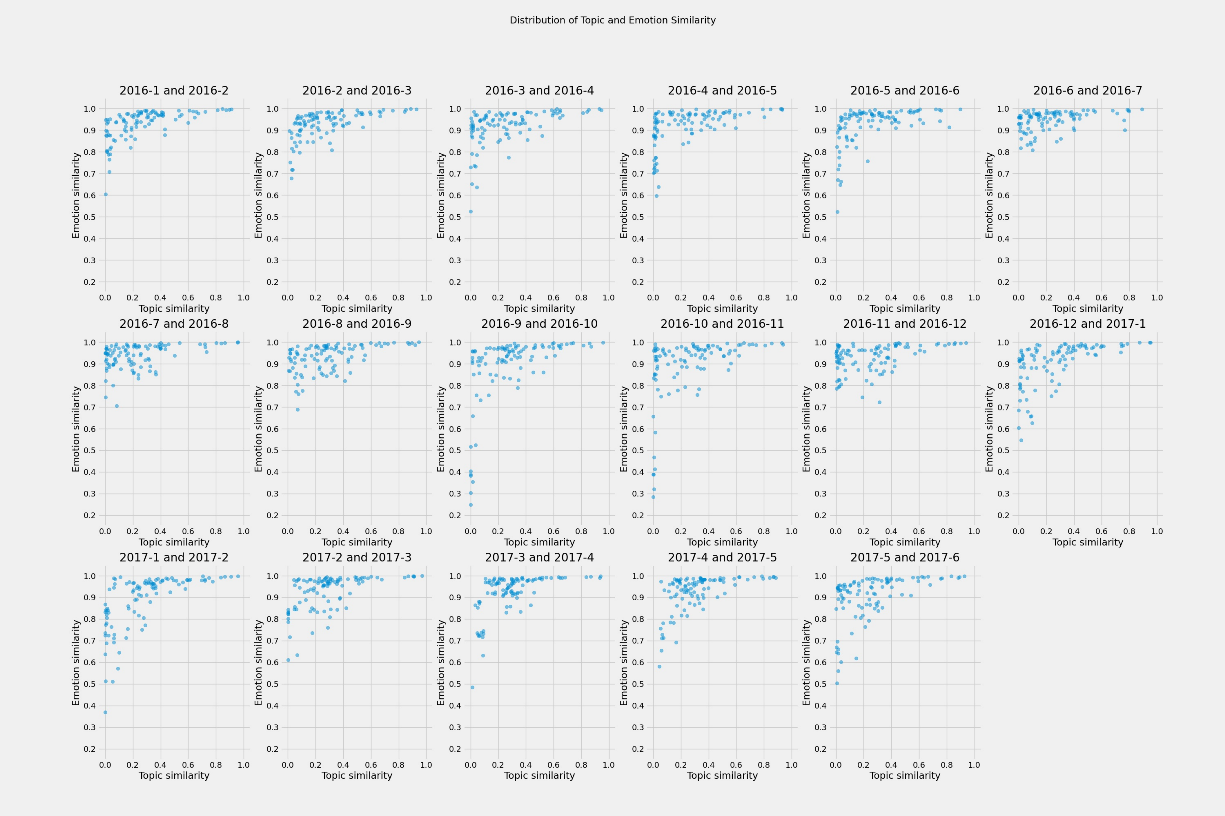

Figure 5: Distribution of topic and emotion similarity

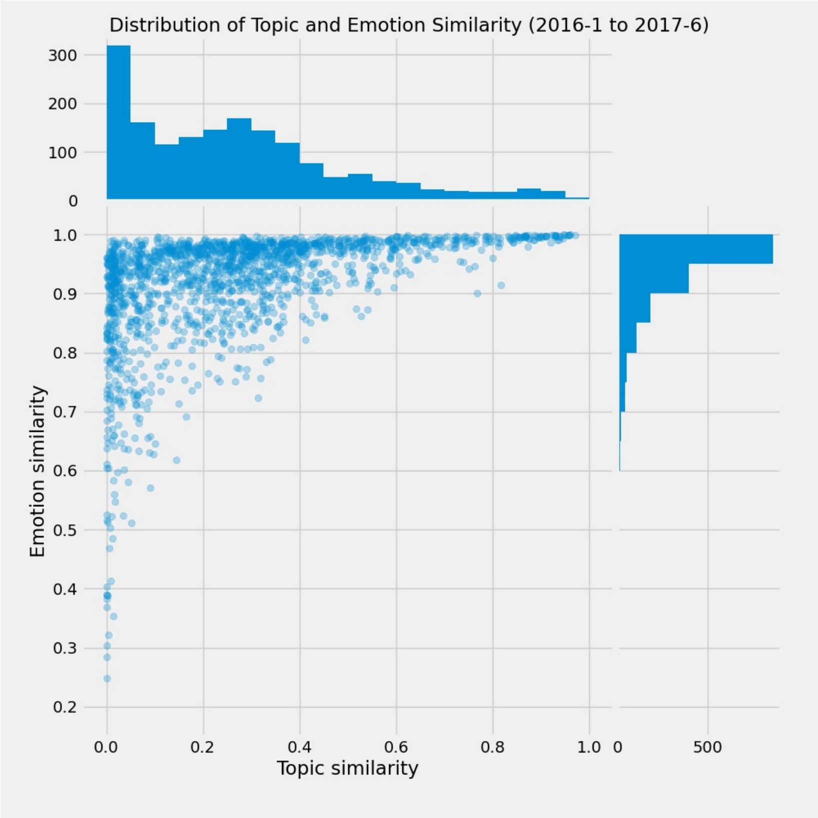

Figure 6: Combined distribution of topic and emotion similarity

Finally, in Figure 7, the evolution differences through the 18 months that we consider are included. They can be observed to stay stable mostly between 0.6 and 0.7, showing that the topic and emotion evolutions are very different. The computed values are shown in Table 1.

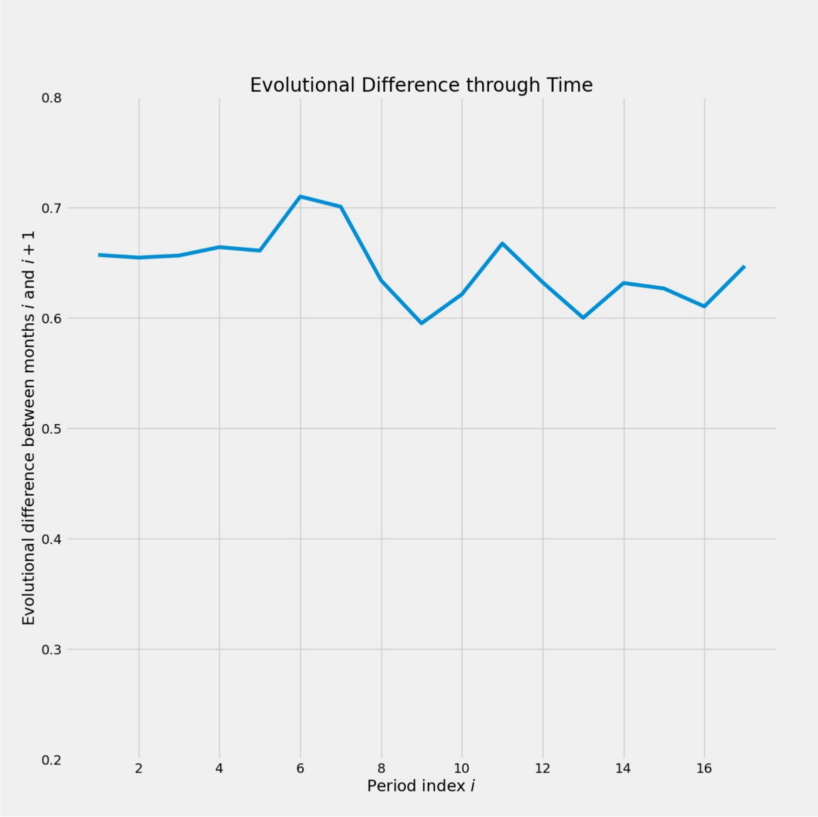

Figure 7: Evolutional difference through time

Period k	Evolutional difference at time k
1	0.656
2	0.655
3	0.656
4	0.664
5	0.661
6	0.710
7	0.701
8	0.634
9	0.595
10	0.621
11	0.667
12	0.632
13	0.600
14	0.632
15	0.627
16	0.610
17	0.647

Table 1: Evolutional difference through time

5 Discussion

Our most important result was that *topics and emotions have different evolutional patterns.* This was reinforced by

- The evolutional difference metric ranging between 0.6 and 0.7 consistently.
- A series of observations which can be explained as follows.

5.1 Emotions in news are unified

The very high similarity in emotion we observed might have been due to the fact that news articles generally employ consistent journalism language, i.e., straightforward and objective. As such, no matter what topics are discussed, they are likely to be all written with a vocabulary that contains little personal emotional interference.

5.2 Each topic is similar to at most one topic in the next month

For any month, each considered topic is highly similar to at most one topic in the following month.

- It is very likely that the highly-similar topic is in fact the same topic, indicating that it is still trending and has not yet died out.
- On the other hand, a topic cannot be similar to several topics. As seen above, each month, there are many available distinct topics even if they overlap a little in terms of words. This gives rise to the sparse topic matrices.

5.3 Low topic similarity may correspond to a large range of emotion similarity

We observed that low topic similarity does not necessarily correspond to low emotion similarity. This can be explained when we revisit the formula for induced emotion:

$$ed(t, e) = \frac{\sum_{w \in V, w(e)=1} t(w)}{\sum_{e' \in E} \sum_{w \in V, w(e')=1} t(w)} . \tag{5.1}$$

Here, as emotions are computed in aggregation from words of a topics, it is natural that emotions of different topics can be similar as long as their words contain the same emotions. In turn, that words have the same emotions does not imply that they are the same: the mapping from words to emotion is highly non-injective in the NRC lexicon. This explains our observation.

In contrast, similar topics have similar emotions. This is also explainable using the definition of induced emotions: it can be directly checked that $ed(t, e) \sim ed(t', e)$ if $t \sim t$.

6 Conclusions

In this work, we examined the evolution of topics and emotions in news to see whether they are related. To this end, we defined novel concepts describing the evolution of topics and emotions as well as their difference, namely the temporal similarity tensor and the evolutional difference. We concluded that *topics and emotions have different evolution patterns*. This is supported by the *high evolutional difference metric*, and a number of explainable observations based on the temporal similarity tensor.

Our research has its limitations, however:

- The main limitation is *algorithmic confounding*; each newspaper has its own internal mechanism that affects which articles are featured in the home-page headlines and in RSS feeds. As such, a fair representation of news articles and their topics is not ensured.
- Since all the publications were American, the data is Western-centric and local in its portrayal of news.
- The popularity of the topics is not considered by the temporal similarity tensors, so we cannot see how the popularity of a topic correlates with different evolution patterns.
- There might be a more optimal number of topics, rather than 10, that would give more quality topic descriptions. As we can observe in topic 8 of January 2017, the words `like, make, just, people` do not really point to any meaningful theme.
- As Latent Dirichlet Allocation is a probabilistic model, the result may change as we fit the data several times. It would be better to observe whether the results are consistent.
- The data representation that Latent Dirichlet Allocation uses does not include how words are used together in the sentence. This hinders the understanding of context, hence lowering the quality of topics. We can improve this by considering n-grams, i.e., phrases of n consecutive words at a time, rather than just individual words.

Investigating the relation between a topic's popularity and its evolution could be an interesting topic of further research. In addition, to compare how journalism differs across the world, newspaper articles from different countries could be analysed and their results compared and contrasted. Since newspapers articles do not display much varied emotion (as we discussed above), performing a similar analysis on a more opinionated sources of news, such as Reddit, Twitter or online news blogs, could yield more diverse results.

References

[1] Alexandra Balahur et al. "Sentiment Analysis in the News". en. In: (Sept. 2013). URL: https://arxiv.org/abs/1309.6202v1 (visited on 12/17/2021).

[2] David Blei, Andrew Ng, and Michael Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1002.

[3] Baitong Chen et al. "Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval". en. In: *Journal of Informetrics* 11.4 (Nov. 2017), pp. 1175–1189. ISSN: 1751-1577. DOI: 10.1016/j.joi.2017.10.003. URL: https://www.sciencedirect.com/science/article/pii/S1751157717300536 (visited on 12/17/2021).

[4] Tomohiro Fukuhara, Hiroshi Nakagawa, and Toyoaki Nishida. “Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events”. In: 2007.

[5] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. “Large-Scale Sentiment Analysis for News and Blogs”. In: 2007.

[6] Erin Hea-Jin Kim et al. “Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news”. en. In: *Journal of Information Science* 42.6 (Dec. 2016). Publisher: SAGE Publications Ltd, pp. 763–781. ISSN: 0165-5515. DOI: 10.1177/0165551515608733. URL: https://doi.org/10.1177/0165551515608733 (visited on 12/17/2021).

[7] Ali Al-Laith and Muhammad Shahbaz. “Tracking sentiment towards news entities from Arabic news on social media”. en. In: *Future Generation Computer Systems* 118 (May 2021), pp. 467–484. ISSN: 0167-739X. DOI: 10.1016/j.future.2021.01.015. URL: https://www.sciencedirect.com/science/article/pii/S0167739X2100025X (visited on 12/17/2021).

[8] Peng Liu, Jon Atle Gulla, and Lemei Zhang. “Dynamic Topic-Based Sentiment Analysis of Large-Scale Online News”. en. In: *Web Information Systems Engineering – WISE 2016*. Ed. by Wojciech Cellary et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 3–18. ISBN: 978-3-319-48743-4. DOI: 10.1007/978-3-319-48743-4_1.

[9] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. “NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets”. In: *arXiv:1308.6242 [cs]* (Aug. 2013). arXiv: 1308.6242. URL: http://arxiv.org/abs/1308.6242 (visited on 12/17/2021).

[10] Neil O’Hare et al. “Topic-dependent sentiment analysis of financial blogs”. In: *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*. TSA ’09. New York, NY, USA: Association for Computing Machinery, Nov. 2009, pp. 9–16. ISBN: 978-1-60558-805-6. DOI: 10.1145/1651461.1651464. URL: https://doi.org/10.1145/1651461.1651464 (visited on 12/17/2021).

[11] Andrew Thompson. *All the news*. en. 2017. URL: https://kaggle.com/snapcrack/all-the-news (visited on 12/17/2021).