人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第32回)
SIG-AM-32-08

# ニュースの内容とコメントの類似度が投稿するコメントに与える影響の分析

# Effect Analysis of the Similarity of News and Posted Comments on New Comments

西原陽子 [1*]　Sheng Wang[2]　Junjie Shan[3]
Yoko Nishihara[1]　Sheng Wang[2]　Junjie Shan[3]

[1] 立命館大学情報理工学部
[1] College of Information Science and Engineering, Ritsumeikan University
[2] 立命館大学大学院情報理工学研究科
[2] Graduate School of Information Science and Engineering, Ritsumeikan University
[3] 立命館グローバルイノベーション研究機構
[3] Ritsumeikan Global and Innovation Research Organization

**Abstract:** Comments on news sites play an important role in understanding the news. On the other hand, many comments are not necessarily highly relevant to the news. Comments with low relevance may hinder the understanding of the news. We hypothesized that new comments would be more relevant to the news if the comments that had already been posted showed high relevance to the news. In this paper, we conducted an experiment in which we evaluated the relevance between news and comments in terms of similarity, and clarified the effect of the similarity of comments on newly posted comments. In the experiment, we used 20 articles in four categories from Yahoo! News. We used Japanese-RoBERTa to obtain embedding vectors for the news articles and their comments. The similarity between the news embedding and each comment's embedding was calculated using the cosine similarity. The participants were presented with comments with high similarity and comments with low similarity, and asked to write a new comment after reading the news and the comments. The experimental results showed that the similarity of the new comment was affected by the similarity between the news and the presented comments.

## 1　はじめに

オンラインニュースサイトやソーシャルネットワーキングサイト（SNS）では，ニュースや個人の投稿に対しコメントを書き込むことが一般的に行われている．コメントには個人の意見が含まれており，ニュースや投稿に対し他者が持つ意見を容易に知ることができるようになっており，ニュースの理解に役立っている．Stroudらの研究により，米国に住む人の 55.3%の人がニュースに対してコメントを読み書きする，また 24.6%の人がコメントは書かないが読むことが明らかになっている [5]．ニュースの理解において，コメントの重要性が高くなっている．

一方で書き込まれるコメントの中には，ニュースとの関連が高いものもあれば，高くないものも含まれている．関連が高いコメントはニュースに対する示唆を与え，より深い考察や議論へと発展させる可能性を持つ．反対に関連が高くないコメントはニュースを読む人を混乱させ，他の人の意見を誤解させるなどの影響を与える可能性がある．関連が高いコメントと高くないコメントが混在していると，ニュースやコメントを読む上で支障が生じると考えられる．

フィルタリングやランキングにより，関連が高くないコメントを上位に表示しない手法が存在するが，ニュースに対して書かれるコメントは，関連が高くないものは少ないほうがニュースの理解には望ましい．関連が高くないコメントがあまり書かれないようにするにはどうすればよいか．著者らは，既に書かれたコメント欄に関連が高いコメントが表示されていれば，新しく書かれるコメントも関連が高くなるのではないかと考えた．そこで本論文では，ニュースとコメントの関連

*立命館大学情報理工学部，滋賀県草津市野路東 1-1-1，
nisihara@fc.ritsumei.ac.jp

度を類似度で評価し，提示される既存コメントの類似度が新たに投稿するコメントに与える影響を実験により明らかにする．

### 1.1 既存研究

関連が高いコメントと高くないコメントを分類する研究は既存研究で行われている．例えば，Mozafari らは SNS 上のニュース記事に対するコメントの類似度を評価することで，関連が高いコメントと関高くないコメントを分類する手法を提案している [4]．手法を用いた評価実験の結果から，関連が高いコメントは客観的でニュース内容に関することが書かれている事が多く，高くないコメントは主観的でニュース内容に関することが書かれていないことが多いことが明らかになっている．また，Kolhatkar らはニュースに対するコメントを建設的なものとそれ以外に分類する手法を提案している [3]．彼らの研究で建設的なコメントは，記事と関連し，かつ感情的な反応を誘発しない明確な議論となるものである．手法を用いた評価実験において，コメントの建設性と毒性（汚い言葉遣い・攻撃的なコメント・ヘイトスピーチなど）との関連も分析しており，分析の結果，建設性が高いからといって毒性が低いわけではなく，建設的なコメントと非建設的なコメントの間で毒性の差は見られないことを明らかにしている．このように，関連が高いコメントと高くないコメントを分類する研究は存在するが，関連が高いコメントや低いコメントを提示したときに，新たに書かれるコメントのニュースとの関連がどうなるかを調べた研究は少ない．本論文では同一ニュースに対し，類似度が異なるコメント群を提示することで，新たに書かれるコメントの類似度がどのようになるかを実験で明らかにする．

## 2 分析手法

本論文では，ニュースとコメントの類似度を算出し，類似度が高いコメント群と低いコメント群を作成して実験を行い，実験結果を分析する．はじめに，ニュースとコメントの類似度を算出する方法について説明する．

図 1 に，類似度が高いコメント群と低いコメント群を作成する流れを示す．ニュースの記事とニュースに付与されたコメント集合を入力すると，それぞれの埋め込み表現を獲得する．ニュース記事の埋め込み表現と各コメントの埋め込み表現の類似度を算出し，類似度が高い順にコメントを並び替える．コメントの類似度の閾値以上のものを類似度が高いコメント群，閾値以下のものを類似度が低いコメント群とする．

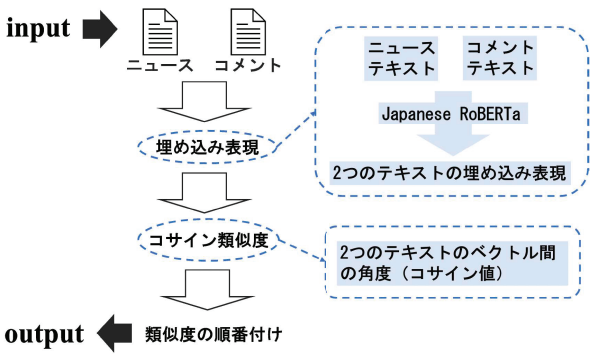


図 1: 類似度が高いコメント群と低いコメント群を作成する流れ．

### 2.1 埋め込み表現を獲得するモデルの選定

本研究ではニュース記事とコメントの関連度を類似度で評価する．このため関連度を類似度で評価できる適切なモデルを選定するために予備実験を行った．

実験でははじめに，ニュースとコメントの関連度について被験者アンケートを取った．複数の埋め込み表現のモデルを用意し，ニュースとコメントの埋め込み表現を獲得し，類似度を算出した．最後に関連度と類似度の決定係数を算出した．

被験者アンケートで使用したニュースは Yahoo!ニュースに掲載された 3 つのニュースであった．ニュースのジャンルは国際であった．各ニュースに付与されたコメントを無作為に 90 件ずつ抽出してアンケートに用いた．被験者は全部で 30 名で，20 代の大学生であった．被験者にはニュースとコメントを読んでもらい，コメントのニュースとの関連の有無の程度を 4 段階で評価してもらった．4 段階評価の 0 が全く関連しない，3 が強く関連するとした．1 つのコメントを 5 名の被験者に評価してもらい，5 名の評価値の平均を正解の関連度とした．

用意した埋め込み表現のモデルは 2 つあり，日本語の Wikipedia で訓練した Doc2Vec [2] と Japanese-RoBERTa [1] の 2 つであった．類似度は 2 種類の方法で評価し，コサイン類似度と内積の 2 つであった．

実験結果を説明する．表 1 に各モデルと各類似度計算方法での決定係数を示す．決定係数が最も高いのは，Japanese-RoBERTa とコサイン類似度を用いた場合であった．このため本論文では Japanese-RoBERTa とコサイン類似度でニュースとコメントの関連度を評価する．

表 1: モデルと類似度計算方法ごとの決定係数.

| モデル | 類似度計算方法 | 決定係数 |
|---|---|---|
| Wikipedia-Doc2vec | コサイン | 0.0997 |
| | 内積 | 0.1345 |
| Japanese-RoBERTa | コサイン | 0.1608 |
| | 内積 | 0.0852 |

# 3 実験

同一ニュースに対して類似度の異なるコメント群を提示することにより，新たに投稿されるコメントの類似度がどのようになるかを実験により調べた.

## 3.1 実験目的

以下の 2 点を明らかにすることを目的とした.

1. ニュースに対して新たにコメントが投稿されるとき，既に書かれたコメントの内容を参考にしているか.
2. 既に書かれたコメントの内容を参考にするとき，新たに投稿されるコメントと既に書かれたコメントの類似度は高くなるか.

1. については，新たに書かれるコメントが提示するコメントの影響を受けるかを調べるためには，被験者が提示されたコメントを参考にしているかどうかを知る必要があると考えたため，実験の目的にいれた．2. については，1. が成り立つ場合に新たに書かれるコメントが提示されるコメントの影響をどの程度受けるかを知るため，実験の目的に入れた.

## 3.2 実験手順

実験は以下の手順で行った.

1. 先に示した分析手法を用い，ニュースと各コメントの類似度を算出し，類似度が高い群と低い群に分ける．それぞれの群から 20 件のコメントを抽出する.
2. 実験者はニュースと 20 件のコメントを提示する．被験者は，ニュースについての新たなコメントを書く.
3. 実験者は得られたコメントと提示コメントの類似度を評価する．新しく得られたコメントは，提示コメントの影響を受けるかを評価する．類似度が閾値以上であるコメントを抽出する.
4. 新しく得られたコメントの中から，類似度が閾値以上であるコメントを抽出する．実験者は抽出されたコメントとニュースの類似度を算出する.

手順の詳細を説明する．手順 1 において，用いたニュースは Yahoo!ニュースの「国際」「国内」「エンタメ」「スポーツ」の 4 つのカテゴリに属する，それぞれ 5 件であり，合計 20 件であった．2023 年の秋に行われた実験実施時期から 3ヶ月以内に発表され，既に投稿されたコメントが 500 件以上あるニュースを用いた．ニュースと各コメントの類似度を算出し，類似度が 0.8 以上であったものを類似度が高いとし，類似度が 0.6 以下であったものを類似度が低いとした．類似度がそれぞれ閾値を満たすコメントの中からランダムに 20 件ずつコメントを抽出した．提示されたコメントの平均文字数は約 90 文字であった.

手順 2 において，実験者は Yahoo!クラウドソーシングで実験を行った．実験者は 1 つのニュースと 20 件のコメントを提示し，被験者にニュースとコメントを読んだ上で自分の新たなコメントを書いてくださいと指示した．被験者はニュースについての新たなコメントを書き，5 つのニュースで先の作業を繰り返した．実験に参加した被験者の総数は 600 名であった．ニュースやコメントを読んだことを確認するため，被験者にはニュース内容に関するクイズを解いてもらい，クイズに正解したコメントのみを収集した．さらに，提示コメントと同じコメント，文字数が少ないコメント（20 文字未満），意味のない文字列の羅列となっているコメントも適切に書かれなかったと判断し，分析の対象外とした.

手順 3 において，得られたコメントと提示されたコメントの類似度を評価する．類似度が閾値以上であれば，提示されたコメントを参考にして書かれたと評価する．最後に手順 4 において，書かれたコメントとニュースの類似度を評価する.

## 3.3 実験結果

初めに，分析対象となったコメント数を示す．得られたコメントから分析対象外のコメントを除いた結果，類似度が高いコメントを提示した群では 1,420 件，類似度が低いコメントを提示した群では 1,404 件のコメントが得られた.

続いて，得られたコメントとニュースの類似度を算出したところ，類似度が高いコメントを提示した群では平均 0.609，類似度が低いコメントを提示した群では平均 0.608 となった．2 つの群の間で類似度の差は見られなかった.

新たに得られたコメントと提示したコメントの類似度の平均値を算出し，平均値を閾値としてコメントを

抽出することで提示コメントを参考にして書かれたコメントを抽出した．抽出されたコメントとニュースの類似度を算出したところ，類似度が高いコメントを提示した群では平均 0.694，類似度が低いコメントを提示した群では平均 0.610 となった．提示コメントを参考にしてコメントを書いた場合，類似度が高いコメントを提示する方が書かれるコメントの類似度が高くなった（$p<0.05$）．

### 3.4 考察

初めに得られたコメントとニュースとの類似度について考察する．類似度が高いコメントを提示した群と類似度が低いコメントを提示した群を比較したところ，コメントとニュースの類似度に差は見られなかった．この結果は，被験者はニュースとコメントを読んだ上で自分のコメントを書くように指示されても，既に投稿されたコメントの影響は余り受けない被験者が一定以上いるということを示している．

続いて，提示コメントを参考にして書かれたと思われるコメントとニュースとの類似度について考察する．2 つの群を比較したところ，類似度が高いコメントを提示した群の方が新たに書かれるコメントとニュースの類似度が高くなった．この結果は，自分のコメントを書くときに既に投稿されたコメントを参考にする場合は，ニュースに関連するコメントを書くことが多いということを示している．

以上より，新たな投稿がコメントされるときに，既に投稿されたコメントを参考にできるようにすることで，ニュースと関連するコメントが投稿される可能性が示された．

## 4 おわりに

本論文ではニュースとコメントの関連度を類似度で評価し，提示される既存コメントの類似度が新たに投稿されるコメントに与える影響を明らかにする実験を行った．類似度は Japanese-RoBERTa による埋め込み表現のコサイン類似度で評価し，Yahoo!ニュースの 4 つのカテゴリに属する 20 件の記事を選んで実験に用いた．実験では被験者に類似度の高いコメントと低いコメントを提示し，ニュースとコメントを読んだ上で新たなコメントを書いてもらった．実験の結果，コメントを新たに書く場合に既に書かれたコメントの影響を受けないことが多いと分かった．そして，提示されたコメントを参考にして新たなコメントを書く場合，提示されたコメントのニュースとの類似度に影響を受け，高いコメントを提示された群は新たなコメントもニュースとの類似度が高くなることが分かった．

## 謝辞

本研究の一部は科研費（22H03708）と立命館グローバルイノベーション研究機構の支援を受けて行われました．記して謝意を申し上げます．

## 参考文献

[1] 趙天雨，沢田慶：日本語自然言語処理における事前学習モデルの公開，第 93 回言語・音声理解と対話処理研究会, pp. 169-170 (2021)

[2] Andrew M. Dai and Christopher Olah and Quoc V. Le: Document Embedding with Paragraph Vectors, arXiv, 1507.07998 (2015)

[3] Varada Kolhatkar and Maite Taboada. Constructive Language in News Comments. In Proceedings of the first workshop on abusive language online, pp. 11-17 (2017)

[4] M. Mozafari, R. Farahbakhsh and N. Crespi: Content Similarity Analysis of Written Comments under Posts in Social Media, 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 158-165 (2019)

[5] Natalie Jomini Stroud, Emily Van Duyn, and Cynthia Peacock: News Commenters and News Comment Readers. Engaging News Project, pp. 1-21 (2016)