University *of Ljubljana*
Faculty *of Computer and Information Science*

Engagement Prediction Of News Articles Based On Natural Language Processing Techniques

Šimen Ravnik, Tim Vučina, and Andrej Kronovšek

Abstract

We present our idea for the course project, roughly outline our goals and means to achieve them. We introduce the core databases on which we intend to do most of the processing and briefly go through the related work and analyse what has already been done in the field of predicting News Article Engagement. We present preliminary results and problems we have encountered including some feasible solutions and future work ideas.

Keywords
Natural Language Processing, Transformers, Article Engagement Prediction

Advisors: Slavko Žitnik

Introduction

User engagement prediction is an interesting topic because it can help us discover and better understand human psychological traits. Besides the analytical benefits it also has potential for use in real-world applications. Being able to predict how well an article will fare after it is published may help writers and publishers to adapt their writing and marketing strategies in order to improve their engagement scores. For the course project we have set ourselves a challenge of predicting news article engagement based on their content using different natural language processing (NLP) approaches.

We will focus specifically on news articles as they are publicly available and more importantly also have enough representation in the Slovenian language as we intend to run and test our models on Slovenian news articles as well. We will train our models using the English based Kaggle dataset [1] and a Slovenian news page RTV [2] from which we will create our own dataset of articles and their prediction scores. The approaches which we will attempt to incorporate include standard machine learning approaches [3] and deep learning models such as classical neural nets, convolutional neural nets and transformers [4], [5].

Related Work

User engagement prediction of online news articles is a relatively new problem which has not yet been researched in depth. However many studies have been performed on engagement and popularity prediction of other content (e.g. tweets) which is in its essence very similar to our proposed problem. Some studies have for instance analysed the rate of content being dispersed online, more precisely the prediction of the number of retweets on Twitter [6] or the prediction of views on YouTube [7], [8]. Some have also tackled similar problems on news article data, however this research was mostly focused on categorising news articles with different topics using semantic analysis [9] and sometimes also predicting the number of comments a certain article will receive [9], [10].

Web content popularity prediction can be performed in two ways: before the content is posted and right after the content is posted. When predicting popularity of content that has already been posted, we use early stage measurements to quickly estimate the engagement [11], [8], but this is often a more difficult task since the engagement in early stages is highly dependent on the content itself. For this reason the predictions are more often made before the content is posted [12]. The authors of [12] used the number of times a given article was posted/shared on twitter alongside with some contextual functions to predict its popularity. The number of times an article is mentioned on twitter is usually less effective than the number of views (which is often available only to the content owner), but can still be used as a popularity measure.

The mentioned studies addresses the problem of predicting the engagement of text content as one of regression [13], classification [10] or clustering problems [7]. Many different approaches were used in these studies, but they can mostly be separated into two categories: time series engagement prediction and content based prediction. The content based

approaches use features like sentiment analysis [14], emotional expressions [15], subjective language [12], people identities [12], and current hot content, which are all considered to be correlated with the engagement. Due to these correlations some researchers have been able to effectively deal with similar problems without the use of large or complex neural networks. Authors of [3] have tackled the problem using a standard K-means approach and have achieved a promising result of 80% accuracy.

Dataset

For the purpose of learning and testing the models on articles written in Slovenian language, we created our own dataset by scrapping one of the most popular Slovenian news sites - RTV [2]. We collected data from more than 10.000 Slovenian articles from multiple categories. The features that were collected are presented in Table 1.

	Feature name	Feature type
Article features	url	*string*
	author	*string*
	datetime_published	*string*
	category	*string*
	title	*string*
	subtitle	*string*
	headline	*string*
	content	*string*
	tags	*list[string]*
Engagement features	total_comments	*int*
	comments	*list[comment]*

Table 1. List of features in our dataset of slovenian articles

User engagement

To measure the engagement of articles we decided to look at the comment count of each article – because this was the only metric that we are able to obtain from the articles. For the purpose of checking if the measure is *good* we took the steps described in the next few sections.

0.1 Scraping 24ur.com

We theorized that multiple news pages publish articles about the same topics with similar news titles. This is why we chose another news website and scraped its titles. We got 150.000 titles we could the correlate our titles from rtvslo.si articles. We chose 24ur.com because it has proven to be the easiest to scrape and has a decent number of comments on the articles.

0.2 Lemmatizer

We have the lemmatized our two dataset titles using **Classla** engine which is fork of Stanza but is focused on the Slavic languages (Slovenian, Croatian, Serbian, etc.). Lemmatization is an essential process when we are calculating textual similarity since the words with the same meaning could be used differently in similar titles. We used lemmatized titles as inputs to our similarity model.

0.3 Similarity model

The core idea for our similarity model is based on the **TF-IDF** measure. For each original article title from rtvslo.si we calculated similarity score for a window of article titles from 24ur.com. The window was set to 4 days around the time the publishing of the article. The we picked the one title that was the most similar to the original article regardless of the similarity score. Later on, we set a meaningful threshold to really extract the articles that are discribing the same content.

0.4 Scraping 24ur.com – again

Now that we determine similar articles for our original RTVSLO dataset, we needed to retrieve also the number of comments that the article received on 24UR. With the new mini dataset of the article titles that could be correlated we have scraped 24ur.com again – this time also the number of comments.

0.5 Correlating number of comments

Our goal for this process was to determine, whether our engagement measure (number of comments of article) is indeed the right one for our predictions. Therefore we wanted to prove this using two different news articles sites and compare them between each other. More precisely we wanted to calculate the correlation between number of comments of article from each of the web sites. If number of comments from RTVSLO would correlate to the number of comments from 24UR, this would mean that our measure is representative and we can use it for model training.

0.6 Results

As already mentioned, retrieved number of comments from both RTVSLO and 24UR to see if there exists any correlation. We can plot these two variables on a 2D plane where we could quickly see the relationship between them. We can see the result in Figure 1.

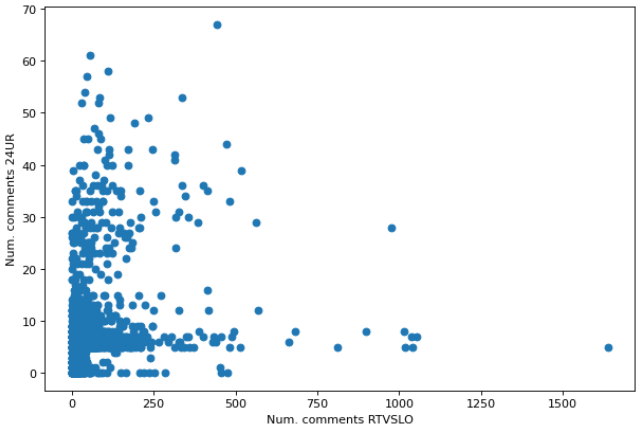

Figure 1. Correlation between number of articles from 24UR and RTVSLO.

From the Figure 1 we can see that the two variables are not exactly correlated. Our assumption was that if an article received a lot of comments on one page, the same would happen on the other page. But we can see that this is not the case. Moreover we can see exactly opposite situation here. Meaning that if an article received many comments on one page, there weren't many on the other page. Our explanation for this is that people are following both of these pages and if the conversation evolves on one page, this would not happend on the other page. Therefore with this results we cannot determine whether our user engagement measure is a proper one for our goal.

With this realization, we quickly see that we need a different user engagement measure. And the optimal measure for user engagement is ofcourse number of views of certain article. But this information is not available on neither of the pages. So we contacted both RTVSLO and 24UR if they are willing to share this information with us, but we are still awaiting the answer.

Looking again at Figure 1 we can see two interesting horizontal lines at zero and at around 6 comments. This is interesting since the number of comments should be randomly distributed. Therefore we plotted the histogram for both RTVSLO and 24UR.

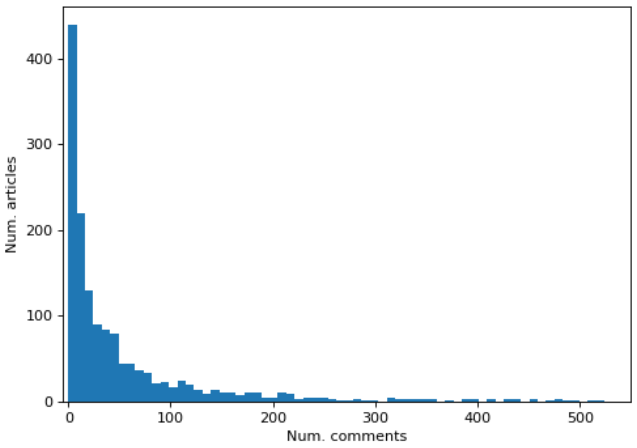

Figure 2. Histogram for number of comments at RTVSLO.

We can see an expected distribution on the Figure 2, where most of the comments are around zero and the number of comments exponentially decreases. But we can see an interesting distribution in the histogram for 24UR seen in Figure 3. We can see that the most comments here are not around zero, but we have a spike at around 6 comments which is interesting. One can argue that this is not genuine distribution, but we cannot be 100% sure.

0.7 Future work

We have contacted rtvslo.si in hopes we could get some sort of data about views of articles and base our prediction model on that.

When analysing the number of comments we have also

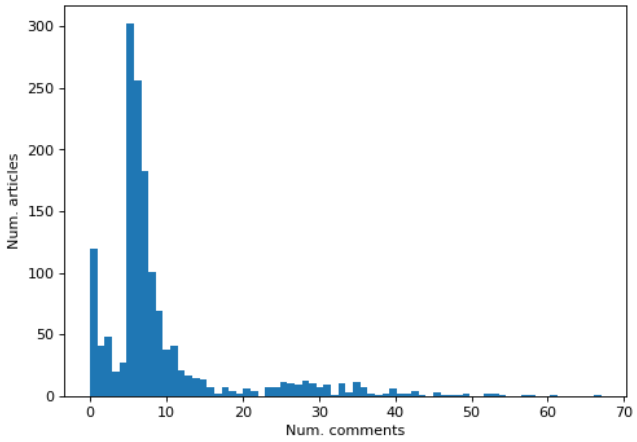

Figure 3. Histogram for number of comments at 24UR.

seen some different distributions of numbers of the correlated graphs. We can see that on Figure 4 where we compare the share of comments in each category for the RTVSLO and 24UR datasets. Where this plots differ are mostly categories: *slovenia*, *health*, *science* and *sport*. We could analyse this further by including more news sites and maybe draw some conclusions on the type of persons that would look at a typical article for each news network.

For the analysis we took articles that had a similarity score of at least 0.4 and this pie plots are essentially done on articles that are very similar. In fact some of the articles that we have actually then checked by reading them, were written almost word for word. This makes us believe that the article was translated from the same original article which was published by other news website. We would like to maybe be able to find the original article for articles that we find to be very similar in content.

References

[1] *Internet news data with readers engagement*. en. URL: https://kaggle.com/szymonjanowski/internet-articles-data-with-users-engagement (visited on 03/17/2022).

[2] *RTVSLO.si*. sl. URL: https://www.rtvslo.si (visited on 03/17/2022).

[3] Ameesha Mittal et al. "User Engagement Prediction Using Tweets". In: July 2018, pp. 802–808. ISBN: 978-3-319-95932-0. DOI: 10.1007/978-3-319-95933-7_88.

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv:1810.04805 [cs]* (May 2019). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 03/17/2022).

[5] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020,

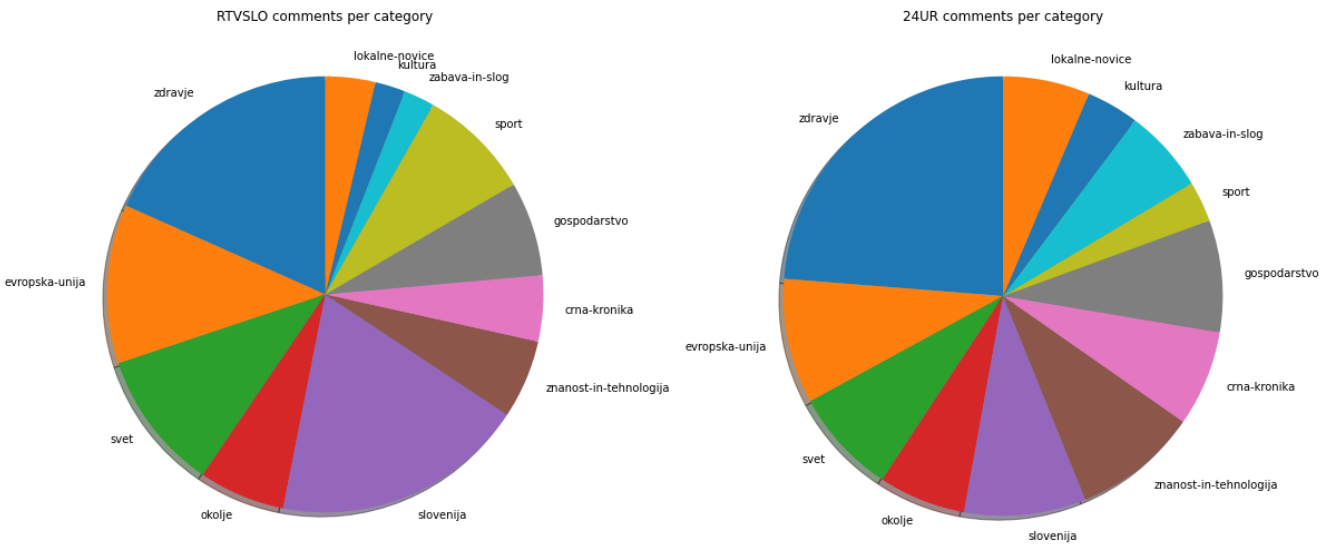

Figure 4. Which articles are commented the most?

pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6 (visited on 03/17/2022).

[6] Tauhid Zaman, Emily B. Fox, and Eric T. Bradlow. "A Bayesian approach for predicting the popularity of tweets". In: *The Annals of Applied Statistics* 8.3 (Sept. 2014). arXiv: 1304.6777. ISSN: 1932-6157. DOI: 10.1214/14-AOAS741. URL: http://arxiv.org/abs/1304.6777 (visited on 03/17/2022).

[7] Gonca Gursun, Mark Crovella, and Ibrahim Matta. "Describing and Forecasting Video Access Patterns". In: *In Proc. of IEEE INFOCOM '11 Mini-Conference*, pp. 11–15.

[8] Gabor Szabo and Bernardo A. Huberman. "Predicting the popularity of online content". In: *Communications of the ACM* 53.8 (Aug. 2010), pp. 80–88. ISSN: 0001-0782. DOI: 10.1145/1787234.1787254. URL: https://doi.org/10.1145/1787234.1787254 (visited on 03/17/2022).

[9] Alexandru-Florin Tatar et al. "Ranking News Articles Based on Popularity Prediction". In: *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (2012). DOI: 10.1109/ASONAM.2012.28.

[10] M. Tsagkias, W. Weerkamp, and M. de Rijke. "Predicting the volume of comments on online news stories". In: *CIKM* (2009). DOI: 10.1145/1645953.1646225.

[11] Luis Marujo et al. "Hourly Traffic Prediction of News Stories". In: *arXiv:1306.4608 [cs]* (June 2013). arXiv: 1306.4608. URL: http://arxiv.org/abs/1306.4608 (visited on 03/18/2022).

[12] Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. "The Pulse of News in Social Media: Forecasting Popularity". In: *arXiv:1202.0332 [physics]* (Feb. 2012). arXiv: 1202.0332. URL: http://arxiv.org/abs/1202.0332 (visited on 03/18/2022).

[13] Jong Gun Lee, Sue Moon, and Kavé Salamatian. "Modeling and Predicting the Popularity of Online Contents with Cox Proportional Hazard Regression Model". In: *Neurocomputing* 76.1 (Jan. 2012). Publisher: Elsevier, pp. 134–145. DOI: 10.1016/j.neucom.2011.04.040. URL: https://hal.archives-ouvertes.fr/hal-00623712 (visited on 03/18/2022).

[14] Julio Reis et al. "Breaking the News: First Impressions Matter on Online News". en. In: (Mar. 2015). DOI: 10.48550/arXiv.1503.07921. URL: https://arxiv.org/abs/1503.07921v2 (visited on 03/18/2022).

[15] Jonah Berger and Katherine L. Milkman. "What Makes Online Content Viral?" en. In: *Journal of Marketing Research* 49.2 (Apr. 2012). Publisher: SAGE Publications Inc, pp. 192–205. ISSN: 0022-2437. DOI: 10.1509/jmr.10.0353. URL: https://doi.org/10.1509/jmr.10.0353 (visited on 03/18/2022).