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.

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 iden-

tities [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

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, 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).