Shallow reading with Deep Learning: Predicting popularity of online content using only its title

Wojciech Stokowiec[1,3], Tomasz Trzciński[2,3], Krzysztof Wołk[1],
Krzysztof Marasek[1] and Przemysław Rokita[2]

[1] Polish-Japanese Academy of Information Technology, Warsaw, Poland,
`wojciech.stokowiec@pjwstk.edu.pl`,
[2] Warsaw University of Technology, Warsaw, Poland
`t.trzcinski@ii.pw.edu.pl`, `pro@ii.pw.edu.pl`
[3] Tooploox, Poland

Abstract. With the ever decreasing attention span of contemporary Internet users, the title of online content (such as a news article or video) can be a major factor in determining its popularity. To take advantage of this phenomenon, we propose a new method based on a bidirectional Long Short-Term Memory (LSTM) neural network designed to predict the popularity of online content using only its title. We evaluate the proposed architecture on two distinct datasets of news articles and news videos distributed in social media that contain over 40'000 samples in total. On those datasets, our approach improves the performance over traditional shallow approaches by a margin of 15%. Additionally, we show that using pre-trained word vectors in the embedding layer improves the results of LSTM models, especially when the training set is small. To our knowledge, this is the first attempt of applying popularity prediction using only textual information from the title.

Introduction

The distribution of textual content is typically very fast and catches user attention for only a short period of time [Castillo et al., 2014]. For this reason, proper wording of the article title may play a significant role in determining the future popularity of the article. The reflection of this phenomenon is the proliferation of *click-baits* - short snippets of text whose main purpose is to encourage viewers to click on the link embedded in the snippet. Although detection of click-baits is a separate research topic [Chakraborty et al., 2016], in this paper we address a more general problem of predicting popularity of online content based solely on its title.

Predicting popularity in the Internet is a challenging and non-trivial task due to a multitude of factors impacting the distribution of the information: external context, social network of the publishing party, relevance of the video to the final user, etc. This topic has therefore attracted a lot of attention from the research

community [Szabo and Huberman, 2010, Pinto et al., 2013, Castillo et al., 2014, Ramisa et al., 2016].

In this paper we propose a method for online content popularity prediction based on a bidirectional recurrent neural network called BiLSTM. This work is inspired by recent successful applications of deep neural networks in many natural language processing problems [Collobert et al., 2011, Zhang and LeCun, 2015]. Our method attempts to model complex relationships between the title of an article and its popularity using novel deep network architecture that, in contrast to the previous approaches, gives highly interpretable results. Last but not least, the proposed BiLSTM method provides a significant performance boost in terms of prediction accuracy over the standard shallow approach, while outperforming the current state-of-the-art on two distinct datasets with over 40'000 samples.

To summarize, the contributions presented is this paper are the following:

– Firstly, we propose title-based method for popularity prediction of news articles based on a deep bidirectional LSTM network
– Secondly, we show that using pre-trained word vectors in the embedding layer improves the results of LSTM models.
– Lastly, we evaluate our method on two distinct datasets and show that it outperforms the traditional shallow approaches by a large margin of 15%.

The remainder of this paper is organized in the following manner: first, we review the relevant literature and compare our approach to existing work. Next, we formulate the problem of popularity prediction and propose a model that takes advantage of BiLSTM architecture to address it. Then, we evaluate our model on two datasets using several pre-trained word embeddings and compare it to benchmark models. We conclude this work with discussion on future research paths.

Related work

The ever increasing popularity of the Internet as a virtual space to share content inspired research community to analyze different aspects of online information distribution. Various types of content were analyzed, ranging from textual data, such as Twitter posts [Castillo et al., 2014] or Digg stories [Szabo and Huberman, 2010] to images [Khosla et al., 2014] to videos [Chesire et al., 2001, Pinto et al., 2013, Trzcinski and Rokita, 2015]. Although several similarities were observed across content domains, e.g. log-normal distribution of data popularity [Tsagkias et al., 2010], in this work we focus only on textual content and, more precisely, on the popularity of news articles and its relation to the article's title.

Forecasting popularity of news articles was especially well studied in the context of Twitter - a social media platform designed specifically for sharing textual data [Bandari et al., 2012, Hong et al., 2011]. Not only did the previous works focus on the prediction part, but also on modeling message propagation within the network [Osborne and Lavrenko, 2011]. However, most of the works

were focused on analyzing the social interactions between the users and the characteristics of so-called social graph of users' connections, rather than on the textual features. Contrary to those approaches, in this paper we base our predictions using only textual features of the article title. We also validate our proposed method on one dataset collected using a different social media platform, namely Facebook, and another one created from various news articles [Ramisa et al., 2016].

Recently, several works have touched on the topic of popularity prediction of news article from a multimodal perspective [Ramisa et al., 2016, Chen et al., 2016]. Although in [Ramisa et al., 2016] the authors analyze news articles on a per-modality basis, they do not approach the problem of popularity prediction in a holistic way. To address this shortcoming, [Chen et al., 2016] have proposed a multimodal approach to predicting popularity of short videos shares in social media platform Vine[4] using a model that fuses features related to different modalities. In our work, we focus only on textual features of the article title for the purpose of popularity prediction, as our goal is to empower the journalists to quantitatively assess the quality of the headlines they create before the publication. Nevertheless, we believe that in future research we will extend our method towards multimodal popularity prediction.

Method

In this section we present the bidirectional LSTM model for popularity prediction. We start by formulating the problem and follow up with the description of word embeddings used in our approach. We then present the Long Short-Term Memory network that serves as a backbone for our bidirectional LSTM architecture. We conclude this section with our interpretation of hidden bidirectional states and describe how they can be employed for title introspection.

Problem formulation

We cast the problem of popularity prediction as a binary classification task. We assume our data points contain a string of characters representing article title and a popularity metric, such as number of comments or views. The input of our classification is the character string, while the output is the binary label corresponding to popular or unpopular class. To enable the comparison of the methods on datasets containing content published on different websites and with different audience sizes, we determine that a video is popular if its popularity metric exceeds the median value of the corresponding metric for other points in the set, otherwise - it is labeled as unpopular. The details of the labeling procedure are discussed separately in the Datasets section.

---

[4] https://vine.co/

Text representation

Since the input of our method is textual data, we follow the approach of [Salton et al., 1975] and map the text into a fixed-size vector representation. To this end, we use word embeddings that were successfully applied in other domains. We follow Collobert et al. [2011] and use pre-trained GloVe word vectors [Pennington et al., 2014] to initialize the embedding layer (also known as look-up table).

Bidirectional Long Short-Term Memory Network

Our method for popularity prediction using article's title is inspired by a bidirectional LSTM architecture. The overview of the model can be seen in Fig. 1.
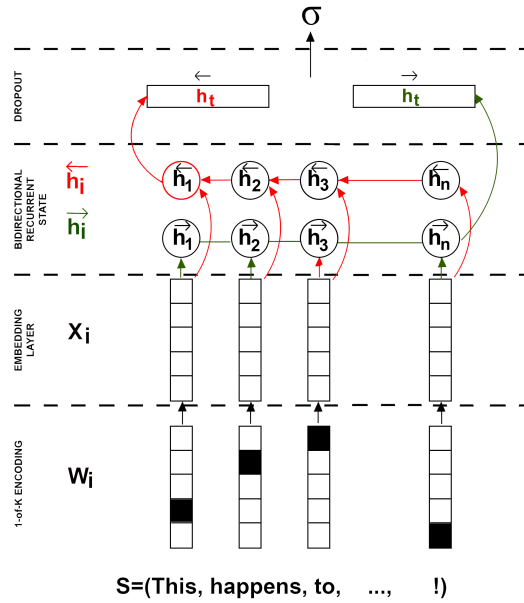

Fig. 1. A bidirectional LSTM architecture with 1-of-K word encoding and embedding layer proposed in this paper.

Let $x_i \in \mathbb{R}^d$ be d-dimensional word vector corresponding to the i-the word in the headline, then a variable length sequence: $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ represents a headline. A recurrent neural network (RNN) processes this sequence by recursively applying a transformation function to the current element of sequence x_t and it's previous hidden internal state h_{t-1} (optionally outputting y_t). At each time step t, the hidden state is updated by:

$$h_t = \sigma(W_h \cdot [h_{t-1}, x_t] + b_h), \qquad (1)$$

where σ is a non-linear activation function. LSTM network [Hochreiter and Schmidhuber, 1997] updates it's internal state differently: at each step t it computes the hidden vector sequence by:

$$
\begin{aligned}
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
\widetilde{c}_t &= \tanh(W_{\widetilde{c}} \cdot [h_{t-1}, x_t] + b_{\widetilde{c}}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \widetilde{c}_t \\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}
\tag{2}
$$

where σ is the *sigmoid* activation function, *tanh* is the hyperbolic tangent function and $\odot$ denotes component-wise multiplication. In our experiments we used 128, 256 for the dimensionality of hidden layer in both LSTM and BiLSTM . The term in equation 2 i_t, is called the input gate and it uses the input word and the past hidden state to determine whether the input is worth remembering or not. The amount of information that is being discarded is controlled by forget gate f_t , while o_t is the output gate that controls the amount of information that leaks from memory cell c_t to the hidden state h_t. In the context of classification, we typically treat the output of the hidden state at the last time step of LSTM as the document representation and feed it to sigmoid layer to perform classification [Zhou et al., 2015].

Due to its sequential nature, a recurrent neural network puts more emphasis on the recent elements. To circumvent this problem [Schuster et al., 1997] introduced a bidirectional RNN in which each training sequence is presented forwards and backwards to two separate recurrent nets, both of which are connected to the same output layer. Therefore, at any time-step we have the whole information about the sequence. This is shown by the following equation:

$$
\begin{aligned}
\overrightarrow{h_t} &= \sigma(\overrightarrow{W_h} \cdot [h_{t-1}, x_t] + \overrightarrow{b_h}) \\
\overleftarrow{h_t} &= \sigma(\overleftarrow{W_h} \cdot [h_{t+1}, x_t] + \overleftarrow{b_h})
\end{aligned}
\tag{3}
$$

In our method, we use the bidirectional LSTM architecture for content popularity prediction using only textual cues. We have to therefore map the neural network outputs from a set of hidden states $(\overrightarrow{h_i}, \overleftarrow{h_i})_{i \in 1...n}$ to classification labels. We evaluated several approaches to this problem, such as max or mean pooling. The initial experiments showed that the highest performance was achieved using late fusion approach, that is by concatenating the last hidden state in forward and backward sequence. The intuition behind this design choice is that the importance of the first few words of the headline is relatively high, as the information contained in $\overleftarrow{h_1}$, i.e. the last item in the backward sequence, is mostly taken from the first word.

Hidden state interpretation

One interesting property of bidirectional RNNs is the fact, that the concatenation of hidden states $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ can be interpreted as a context-dependent vector representation of word w_t. This allows us to introspect a given title and approximate the contribution of each word to the estimated popularity. To that end one can process the headline representation $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ through the bidirectional recurrent network and then retrieve pairs of forward and backwards hidden state $[\overrightarrow{h_t}, \overleftarrow{h_t}]$ for each word x_t. Then, the output of the last fully-connected layer $\sigma([\overrightarrow{h_t}, \overleftarrow{h_t}])$ could be interpreted as context-depended popularity of a word x_t.

Training

In our experiments we minimize the binary cross-entropy loss using Stochastic Gradient Descent on randomly shuffled mini-batches with the Adam optimization algorithm [Kingma and Ba, 2014]. We reduce the learning rate by a factor of 0.2 once learning plateaus. We also employ early stopping strategy, i.e. stopping the training algorithm before convergence based on the values of loss function on the validation set.

Evaluation

In this section, we evaluate our method and compare its performance against the competitive approaches. We use k-fold evaluation protocol with $k = 5$ with random dataset split. We measure the performance using standard accuracy metric which we define as a ratio between correctly classified data samples from test dataset and all test samples.

Datasets

In this section we present two datasets used in our experiments: The NowThis-News dataset, collected for the purpose of this paper, and The BreakingNews dataset [Ramisa et al., 2016], publicly available dataset of news articles.

The NowThisNews dataset contains 4090 posts with associated videos from NowThisNews Facebook page[5] collected between 07/2015 and 07/2016. For each post we collected its title and the number of views of the corresponding video, which we consider our popularity metric. Due to a fairly lengthy data collection process, we decided to normalize our data by first grouping posts according to their publication month and then labeling the posts for which the popularity metric exceeds the median monthly value as popular, the remaining part as unpopular.

---

[5] https://www.facebook.com/NowThisNews

The Breaking News dataset [Ramisa et al., 2016] contains a variety of news-related information such as images, captions, geo-location information and comments which could be used as a proxy for article popularity. The articles in this dataset were collected between January and December 2014. Although we tried to retrieve the entire dataset, we were able to download 38'182 due to the dead links published in the dataset. The retrieved articles were published in main news channels, such as Yahoo News, The Guardian or The Washington Post. Similarly, to The NowThisNews dataset we normalize the data by grouping articles per publisher, and classifying them as popular, when the number of comments exceeds the median value for given publisher.

Baselines

As a first baseline we use Bag-of-Words, a well-known and robust text representations used in various domains [Wang and Manning, 2012], combined with a standard shallow classifier, namely, a Support Vector Machine with linear kernel. We used Chang and Lin [2011] implementation of SVM.

Our second baseline is a deep Convectional Neural Network applied on word embeddings. This baseline represents state-of-the-art method presented in [Ramisa et al., 2016] with minor adjustments to the binary classification task. The architecture of the CNN benchmark we use is the following: the embedding layer transforms one-hot encoded words to their dense vector representations, followed by the convolution layer of 256 filters with width equal to 5 followed by max pooling layer (repeated three times), fully-connected layer with dropout and l_2 regularization and finally, *sigmoid* activation layer. For fair comparison, both baselines were trained using the same training procedure as our method.

Embeddings

As a text embedding in our experiments, we use publicly available GloVe word vectors [Pennington et al., 2014] pre-trained on two datasets: Wikipedia 2014 with Gigaword5 (W+G5) and Common Crawl (CC)[6]. Since their output dimensionality can be modified, we show the results for varying dimensionality sizes. On top of that, we evaluate two training approaches: using static word vectors and fine-tuning them during training phase.

Results

The results of our experiments can be seen in Tab. 1 and 2. Our proposed BiLSTM approach outperforms the competing methods consistently across both datasets. The performance improvement is especially visible for The NowThisNews dataset and reaches over 15% with respect to the shallow architecture in terms of of accuracy. Although the improvement with respect to the other methods based on deep neural network is less evident, the recurrent nature of our method provides much more intuitive interpretation of the results and allow for parsing the contribution of each single word to the overall score.

Table 1. Popularity prediction results on NowThisNews dataset. Our proposed BiL-STM method provides higher performances than the competitors in terms of classification accuracy.

Model	Word Embeddings	fine-tuned	Dim	Accuracy
BoW + SVM				0.5832
CNN	GloVe (W + G5)	no	100	0.6320
	GloVe (W + G5)	yes	100	0.6454
	GloVe (W + G5)	no	200	0.6308
	GloVe (W + G5)	yes	200	0.6479
	GloVe (W + G5)	no	300	0.6247
	GloVe (W + G5)	yes	300	0.6295
	GloVe (CC)	no	300	0.6528
	GloVe (CC)	yes	300	0.6653
LSTM 128	Glove (W + G5)	no	300	0.63081
	Glove (W + G5)	yes	300	0.64792
LSTM 256	Glove (W + G5)	no	300	0.64914
	Glove (W + G5)	yes	300	0.66504
BiLSTM 128	Glove (W + G5)	no	300	0.6552
	Glove (W + G5)	yes	300	0.6479
BiLSTM 256	Glove (W + G5)	no	300	0.6564
	Glove (W + G5)	yes	300	**0.6711**

Table 2. Popularity prediction results on BreakingNews dataset. Our BiLSTM method outperforms the competitors - the performance gain is especially visible with respect to the shallow architecture of BoW + SVM.

Model	Word Embeddings	fine-tuned	Dim	Accuracy
BoW + SVM				0.7300
CNN	GloVe (W + G5)	no	100	0.7353
	GloVe (W + G5)	yes	100	0.7412
	GloVe (W + G5)	no	200	0.7391
	GloVe (W + G5)	yes	200	0.7379
	GloVe (W + G5)	no	300	0.7319
	GloVe (W + G5)	yes	300	0.7416
	GloVe (CC)	no	300	0.7355
	GloVe (CC)	yes	300	0.7394
LSTM 128	Glove (W + G5)	yes	300	0.6694
	Glove (W + G5)	no	300	0.6663
LSTM 256	Glove (W + G5)	yes	300	0.6619
	Glove (W + G5)	no	300	0.6624
BiLSTM 128	Glove (W + G5)	yes	300	0.7167
	Glove (W + G5)	no	300	0.7406
BiLSTM 256	Glove (W + G5)	yes	300	0.7149
	Glove (W + G5)	no	300	**0.7450**

To present how our model works in practice, we show in Tab. 3 a list of 3 headlines from NowThisNews dataset that are scored with the highest probability of belonging to a popular class, as well as 3 headlines with the lowest score. As can be seen, our model correctly detected videos that become viral at the same time assigning low score to content that underperformed. We believe that BiLSTM could be successfully applied in real-life scenarios.

Table 3. Top and bottom 3 headlines from the NowThisNews dataset as predicted by our model and their views 168 hours after publication.

Top 3 headlines	Views
This teen crossed a dangerous highway to play Pokmon Go and then was hit by a car	20'836'692
This dancer dropped her phone in the water but a dolphin had her back	1'887'482
A man shoved a bag of sh*t down this womans pants and was caught on camera	784'588

Bottom 3 headlines	Views
We're recapping some of the biggest stories from last night and this morning	47'803
We're recapping some of the big stories you might have missed	64'740
Violent clashes between protesters and police broke out in Hong Kong	256'357

Conclusions

In this paper we present a novel approach to the problem of online article popularity prediction. To our knowledge, this is the first attempt of predicting the performance of content on social media using only textual information from it's title. We show that our method consistently outperforms benchmark models. Additionally, the proposed method could not only be used to compare competing titles with regard to their estimated probability, but also to gain insights about what constitutes a good title. Future work includes modeling popularity prediction problem with multiple data modalities, such as images or videos. Furthermore, all of the evaluated models function at the word level, which could be problematic due to idiosyncratic nature of social media and Internet content. It is, therefore, worth investigating, whether combining models that operate at the character level to learn and generate vector representation of titles with visual features could improve the overall performance.

---

[6] http://nlp.stanford.edu/projects/glove/

Bibliography

R. Bandari, S. Asur, and B. A. Huberman. The Pulse of News in Social Media: Forecasting Popularity. *CoRR*, abs/1202.0332, 2012.

C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *CSCW*, 2014.

A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. *CoRR*, abs/1610.09786, 2016.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, and T. Chua. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In *ACMMM*, 2016.

M. Chesire, A. Wolman, G. Voelker, and H. M. Levy. Measurement and analysis of a streaming-media workload. In *USITS*, 2001.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398, 2011.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8), 1997.

L. Hong, O. Dan, and B. Davison. Predicting popular messages in twitter. In *Proc. International Conference Companion on World Wide Web*, 2011.

A. Khosla, A. Sarma, and R. Hamid. What makes an image popular? In *WWW*, 2014.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

M. Osborne and V. Lavrenko. V.: Rt to win! predicting message propagation in twitter. In *ICWSM*, 2011.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

H. Pinto, J. Almeida, and M. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *WSDM*, 2013.

A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk. Breakingnews: Article annotation by image and text processing. *CoRR*, abs/1603.07141, 2016.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11), 1975.

M. Schuster, K. K. Paliwal, and A. General. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997.

G. Szabo and B. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8), 2010.

T. Trzcinski and P. Rokita. Predicting popularity of online videos using support vector regression. *CoRR*, abs/1510.06223, 2015.

M. Tsagkias, W. Weerkamp, and M. de Rijke. News comments: Exploring, modeling, and online prediction. In *ECIR*, 2010.

S. Wang and C. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *ACL*, 2012.

X. Zhang and Y. LeCun. Text understanding from scratch. *CoRR*, abs/1502.01710, 2015.

C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau. A C-LSTM neural network for text classification. *CoRR*, abs/1511.08630, 2015.