Artificial News Popularity Detection Based on Telegram Channels in Azerbaijan

Davud RUSTAMOV, Jalal RASULZADE, Shamsaddin HUSEYNOV
Academy of the State Security Service of the Azerbaijan Republic named after Heydar Aliyev, Baku, Azerbaijan
cyber@dtx.gov.az

Abstract

With the exponential growth of digital media, readers face a daunting task of sifting through vast amounts of information to identify important news. This problem is especially critical for media professionals, journalists, and news agencies who need to quickly filter news articles to identify relevant and significant stories. Machine learning models offer a promising solution by automatically classifying news articles based on their significance. In this paper, we propose novel machine learning models for news significance detection, leveraging state-of-the-art deep learning architectures and a dataset of news articles. We evaluate our models using a variety of performance metrics and demonstrate their effectiveness compared to existing methods. Our proposed approach has the potential to significantly improve the efficiency and accuracy of news selection, benefiting both media professionals and readers alike. Furthermore, it can be beneficial to forecast the popularity of fake news and prevent its dissemination in society. Approximately, 2800 Azerbaijani news articles have been collected from telegram and labeled as popular or unpopular according to statistical calculation results. For news popularity detection, application of SVM, Random Forest and Neural network models and their results have been discussed in this paper.

Index terms: machine learning, natural language processing, popularity detection, telegram, text classification

1. Introduction

Artificial Intelligence has become an indispensable tool for analyzing large amounts of data and making predictions in various fields. One of the areas where AI can be applied is news analysis and popularity detection. With the growth of social media platforms, news channels on messaging apps such as Telegram have gained significant attention in recent years. In Azerbaijan, Telegram has become a popular platform for news dissemination, where numerous news channels provide updates on current events. In this research paper, we focus on detecting the popularity of the posts published in news channels on Telegram in Azerbaijan using AI-based techniques. The objective of this study is to develop an artificial intelligence model that can analyze news articles from various channels and classify their level of popularity as either high or low. The proposed model uses natural language processing and machine learning algorithms to identify the relevant features and patterns in the news articles and classify them accordingly. To accomplish this objective, we collect data from various Telegram channels in Azerbaijan, analyze the collected data, and create a machine learning model that accurately classifies news articles based on user engagement metrics. The significance of this research lies in its potential to assist individuals in identifying the relevance and authenticity of the news they consume. In addition, the study results can help news outlets and media organizations to

understand their audience's interests and tailor their news content accordingly. The novelty of this research lies in its application of machine learning algorithms to the context of news popularity detection on Telegram channels in Azerbaijan.

This research paper addresses the issue of the absence of an effective and precise system to classify the popularity of news articles on Telegram channels in Azerbaijan. The absence of such a system poses a significant challenge for individuals, news outlets, and media organizations in identifying the relevance and authenticity of the news they consume and publish, respectively. This research paper seeks to address this problem by developing an AI-based model that can accurately classify the popularity of news on Telegram channels in Azerbaijan.

2. Literature Review

In recent years, there has been a significant increase in the amount of news articles published online. This includes not only news published in official news web pages but also posts broadcasted via social media or messaging platforms. News broadcasted through social media or messaging apps can spread instantly, making them an effective tool for advertising preferred ideas or products to the wider population. Frequently, advertisements are made either by public influencers, web pages or public chats in mobile applications. According to “Similarweb” [1] mobile application ranking for Azerbaijan “Telegram” is the only popular app (in top ten) which can be used not only for the communication but also for disseminating news. Structure of this messenger is analyzed and discussed in “Analysis of telegram, an instant messaging service” [2]. Authors of the paper have developed a crawler for obtaining and future advertisement detection from posts of 185 public channels and groups published in October 2016. The researchers achieved accuracies of 80.5%, 79.9%, and 79.8% from the machine learning models Neural Network, SVM, and Decision Tree, respectively.

Another research was conducted for predicting post promotion on Twitter written in English [3]. In the mentioned paper, the authors aimed to predict the popularity of a tweet, where popularity is the binary variable which is one when the number of retweets exceed given threshold and zero otherwise. To solve this problem the researchers proposed a mathematical model that incorporates syntactic units, temporal information, and neighborhood influence.

Moreover, another paper [4] discusses applications of LightGBM, XGBoost, Logistic Regression, Random Forest and AdaBoost machine learning algorithms for detection of cyberbullying in tweets. About 47k tweets were categorized into six categories based on age, gender, ethnicity, religion and including cyberbullying content. According to the authors best accuracy was obtained in AdaBoost (79.5%) and best one was LightGBM with the performance 85.5%.

Additional worthwhile article [5] analyzes the shortcoming of the existing Twitter interface in fulfilling users' information gathering requests. While Twitter has expanded beyond its conventional role as a social network, most users still use it primarily to connect with their social networks, leading to inaccurate categorization of information. The research introduces Labeled LDA, a partly supervised learning model that maps the content of the Twitter feed into dimensions such as post contents, style, status, and social features. The authors utilize this model to profile people and tweets and demonstrate how it may be used to facilitate information consumption-oriented activities. They report the results of two such projects, demonstrating the efficacy of their technique in enhancing content representation on Twitter.

Upon further investigation of the topic of news sentiment analysis, it is worth considering the research conducted by Balahur et.al. [6], who compares the challenges of opinion mining in news stories with other text genres, such as movie or product reviews. The authors of the paper identified three subtasks that must be addressed: defining the target, separating good and poor news content from positive and negative emotions expressed about the target, and analyzing clearly designated

opinions that are expressly communicated. The report also differentiates three alternative perspectives on newspaper stories which need various techniques to sentiment analysis. The authors carried performed tests to assess the applicability of several sentiment dictionaries for mining views about entities in English language news. They also looked for distinguish between good and bad news, as well as whether topic domain-defining terminology should be disregarded. Although there are few researches in the field of text classification in Azerbaijani language, one of the research projects shows that it is possible to achieve excellent result in this field. The research [7] aimed to determine the sentiment of news articles in Azerbaijani language where researchers obtained 96.79% f1-score by using SVM classifier with TF-IDF vectorization technique. The results revealed that neglecting topic domain-defining terminology was more appropriate in the context of news opinion mining, and techniques that took this into account performed better.

3. Data Collection

In order to collect data from public Telegram channels, we used python programming language and telethon library. Writing python programs using Telethon library allows us to collect Telegram data effortlessly. Furthermore, in order to use the Telethon library, the API ID and API hash were obtained from the official Telegram website, which are mandatory requirements.

There are several Azerbaijani news channels in Telegram which sharing daily news. Four news channels were selected for data collection step based on their popularity among the society. Three months' worth of news from the selected channels were collected, encompassing the period from January 1, 2023, to April 1, 2023. The content of the published news, count of the reactions, replies, and views were collected as a dataset. Table 1 shows basic statistics of the collected data including the name of news channels, subscriber count, and the amount of dataset.

Table 1. Basic statistics of collected data

Channel Name	Subscriber Count	Collected Data
APA	20005	2093
Baku ES	140310	2855
Oxu.Az	23680	1937
Qafqazinfo	21338	2145

Since it is difficult to determine how active subscribers are in different Telegram news channels, we decided to focus on a single news channel in order to remove inconsistencies from the dataset. It has been taken into account that several different news channels might publish the same news with the same content, but have different numbers of reactions, replies, and views, which can cause different labels for the same input while training supervised machine learning models, leading to inconsistencies in the data. Considering the number of subscribers and the amount of collected data, the channel named “Baku Es” was selected for the experiments.

4. Data Pre-processing

Identifying and addressing issues in the dataset were critical aspects of our research as it was crucial to obtain more accurate results while experimenting with various machine learning algorithms. Since our input features were extracted from content of the news, we mainly focused on noisiness of textual data that might affect the training process of classification model.

As the first step of data pre-processing, we removed emojis and unnecessary punctuations from text as they decreased the accuracy of the model. Since our purpose was to detect popularity of the news specifically, we excluded surveys and advertisements from the collected data in order to obtain

a clean news content dataset. Additionally, we analyzed that there were some kinds of news published with very short text which covers the main content in the shared media (image, video) file. As we were interested in analyzing the text of news specifically, we excluded such news items from the dataset. As a final step, all words in the dataset had been converted to the lowercase dispose of difference between same words with different cases. At the end of the data cleaning process, the number of collected data decreased 2855 to 2531.

5. Definition of Annotation

Obtaining a correctly annotated dataset is crucial to achieve better results when training supervised machine learning algorithms. Even through telegram news contains statistical labels such as count of reactions, replies, and views they cannot be characterized as a final annotation metric for popularity detection. All these metrics have been analyzed independently for having correct annotation. Usage of view and reply count were excluded from the final dataset, as the number of views in telegram, which is calculated by scroll number, was not even close to the number of reads. Although replies written by users could be used as proof of interest in a given post, we had to exclude them from the metrics list due to their high sparsity. This sparsity is mainly caused not only by a lack of interest in the given post but also by the channel administration disabling the reply section of published news.

As a result, total number of reactions, which was calculated as a sum of all reactions, was normalized using zero-one technique and divided into two categories: popular and unpopular. News with a normalized reaction value greater than the given threshold, which is the mean of the normalized reaction count of all news in our dataset, were labeled as popular. All other news was labeled as unpopular. Finally, the distribution of classes in dataset is shown in Fig. 1.

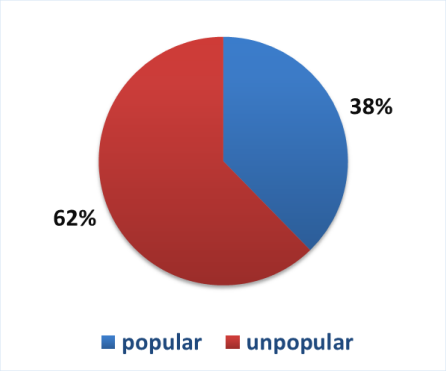

Fig. 1. The distribution of classes

6. Feature Extraction

To enable machine learning algorithms to analyze textual data, we extracted numerical representations of the texts as features using three vectorization methods: count-based, TF-IDF (term-frequency and inverse-document-frequency), and word tokenization. The count-based approach counts the occurrences of each word in a document and creates a vector with these counts. In contrast, the TF-IDF method considers the relative importance of a word within a document and across a dataset [8], taking into account how often it appears in all documents in the dataset. Eventually TF-IDF vectorizer gives more importance to words that are less frequent and less importance to words which are more frequent in a dataset. Moreover, the word tokenization method which translates textual data into sequence of numbers was used as an encoder for neural network classifier. All mentioned features were used in the experiments are described in the next section (Section 7).

7. Methodology

To create a popularity detection model for Azerbaijan news, three different classification algorithms were tested: SVM (Support Vector Machines), Radom Forest, and LSTM (Long short-term memory) based Neural Network.

First classifier SVM have obtained best results with linear kernel and regularization parameter equal to one. Additionally, Random Forest classifier have trained with number of estimators equal to hundred and minimum sample split equal to two. For both of the mentioned models TF-IDF and count vectorization techniques were applied. The last model was created using neural networks which consists of embedding, bidirectional LSTM and dense layers with total number or trainable parameters equal to 138k. The results of all the described models are presented in Table 2.

Table 2. Results

Classifier	Features	Precision	Recall	F1-score	Accuracy
SVM	Count Vectorizer	0.65	**0.62**	0.63	0.72
	Tf-idf Vectorizer	0.71	0.58	**0.64**	**0.75**
Random Forest	Count Vectorizer	**0.81**	0.37	0.51	0.72
	Tf-idf Vectorizer	0.79	0.38	0.51	0.72
Neural Network	Word Tokenization	0.71	0.48	0.58	0.73

8. Conclusion

In conclusion, during the research period we have analyzed several different Azerbaijani news channels in Telegram. As a result, most popular channel, Baku ES, was chosen for experiments and training supervised machine learning algorithms. Having compared the accuracy (Table 2) of all the mentioned the algorithms, we have observed that SVM with TF-IDF vectorizer slightly outperforms with 0.64 f1-score. Furthermore, it was also noticed that both vectorizers showed similar results in Random Forest. The results of the neural network were better than Random Forest but worse than SVM. This can be explained by having features that cannot represent grammatical structure of agglutinative Azerbaijan language well enough. In comparison with analytic languages (Example: English), in agglutinative languages the same word can have different written form depending on its position in the sentence. As an example, we can mention that in Azerbaijani language a single noun may have up to 400 different forms, for the verb this number is even higher (about 600) [9]. Therefore, for obtaining better results and covering specifications of the language it is necessary to user advanced tokenization methods.

As a future work we are planning to increase size of the dataset, test different feature extraction methods (Example: word2vec, BERT), additional features (Example: category and sentiment of the news), and apply different advanced neural network architectures.

References

[1]. A. Hochman, "similarweb," April 2023. [Online]. Available: https://www.similarweb.com/apps/top/google/store-rank/az/all/top-free/. [Accessed 20 April 2023].

[2]. A. Dargahi Nobari, N. Reshadatmand and M. Neshati, "Analysis of telegram, an instant messaging service," Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017.

[3]. C. Xiao, C. Liu, Y. Ma, Z. Li and X. Luo, "Time sensitivity-based popularity prediction for online promotion on Twitter," Information Sciences, vol. 525, pp. 82-92, 2020.

[4]. M. I. Mahmud, M. Mamun and A. Abdelgawad, "A deep analysis of textual features based cyberbullying detection using machine learning," 2022 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT), 2022.

[5]. D. Ramage, S. Dumais and D. Liebling, "Characterizing microblogs with topic models," Proceedings of the International AAAI Conference on Web and Social Media, vol. 4, no. 1, pp. 130-137, 2010.

[6]. A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen and J. Belyaeva, "Sentiment Analysis in the News," Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220, 2010.

[7]. S. Mammadli, S. Huseynov, H. Alkaramov, U. Jafarli, U. Suleymanov and S. Rustamov, "Proceedings - Natural Language Processing in a Deep Learning World," Sentiment polarity detection in Azerbaijani social news articles, 2019.

[8]. P.-H. Chen, H. Zafar, M. Galperin-Aizenberg and T. Cook, "Integrating Natural Language Processing and machine learning algorithms to categorize oncologic response in radiology reports," Journal of Digital Imaging, pp. 178-184, 2017.

[9]. "Dilci," [Online]. Available: www.dilci.az. [Accessed 25 04 2023].