COMPUTER RESEARCH AND MODELING
2026 VOL. 18 NO. 2 P. 501–517
DOI: 10.20537/2076-7633-2026-18-2-501-517

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 004.89

Semi-automated detection of controversy in social media content: an approach based on pre-trained models

A. V. Zaida[a], A. O. Savelev[b]

Tomsk Polytechnic University,
30 Lenina prospect, Tomsk, 634050, Russia

E-mail: [a] avz68@tpu.ru, [b] sava@tpu.ru

Received 15.01.2026, after completion — 28.03.2026.
Accepted for publication 02.04.2026.

Detecting controversy in online discussions is critical for managing public relations, as it helps inform various processes from policymaking to business. This work aims to expand approaches to online controversy detection based on the expressed emotions. Controversy was defined as an online content phenomenon of provoking disagreements and conflict. This study builds upon prior semantic methods by analyzing estimates of emotional connotations of messages. Modern language models for emotion recognition and named entity recognition are explored as tools of controversy detection. The outputs of these models were aggregated by entity to estimate the entity's emotional connotation. The emotional divergence score based on the dispersion of emotions was proposed to quantify controversy in user content. Then, entities with sufficiently high emotional divergence relative to the domain of discussions were selected as markers of controversy. A case study of Reddit data related to Sri-Lankan 2022 political crisis was conducted, showing the capabilities of emotional divergence score in controversy detection. A total of two datasets were collected with different methodologies: one aimed at collecting earlier messages and another aimed at collecting more recent ones. The collected data contained discussions of policy, public figures, organizations and locations tied to the crisis. When measured on manually annotated data samples, the proposed method achieved a recall value of 0.705 and a precision value close to 0.496 for the first dataset, while recall of 0.716 and precision of 0.436 were recorded for the second dataset. The main factors that limit the precision were found to be the quality of underlying models and false positives: highly discussed non-controversial markers. Lastly, it was identified that a study of regular emotional distribution of social media content may be helpful for improving controversy detection quality.

Keywords: controversy detection, social media, natural language processing, sentiment analysis, named entities recognition

Citation: *Computer Research and Modeling*, 2026, vol. 18, no. 2, pp. 501–517.

This study was supported by the Russian Science Foundation (RSF) under Project No. 25-28-01153 "Modeling mechanisms of social contagion in the online radicalization process".

КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2026 Т. 18 № 2 С. 501–517
DOI: 10.20537/2076-7633-2026-18-2-501-517

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 004.89

Автоматизированное выявление противоречивости в контенте социальных медиа: подход на основе предварительно обученных моделей

А. В. Зайда[a], А. О. Савельев[b]

Томский политехнический университет,
Россия, 634050, г. Томск, пр. Ленина, д. 30

E-mail: [a] avz68@tpu.ru, [b] sava@tpu.ru

Получено 15.01.2026, после доработки — 28.03.2026.
Принято к публикации 02.04.2026.

Обнаружение противоречивости в онлайн-дискурсе имеет важное значение для управления связями с общественностью, что позволяет информировать различные процессы от законотворчества до предпринимательства. В данной работе предлагается подход к обнаружению противоречивости в онлайн-контенте на основе анализа выражаемых эмоций. Противоречивость онлайн-контента определяется как феномен провоцирования разногласий и конфликтов в обсуждениях. Данная работа развивает предыдущие семантические методы, анализируя численные оценки именно эмоционального окраса сообщений. В качестве инструментов обнаружения противоречивости рассматриваются современные языковые модели для распознавания эмоций и распознавания именованных сущностей. Результаты работы этих моделей были агрегированы по сущностям для оценки их эмоциональной коннотации. Был предложен показатель эмоциональной дивергенции, основанный на дисперсии эмоций, для количественной оценки противоречивости контента. Затем сущности с достаточно высокой эмоциональной дивергенцией по отношению к специфике коммуникаций в рамках сообщества были отобраны в качестве маркеров противоречивости. Проведены эксперименты на данных Reddit, связанных с политическим кризисом в Шри-Ланке 2022 года, которые подтверждают возможность показателя эмоциональной дивергенции обнаруживать противоречивость. Всего было собрано два набора данных с использованием различных методологий: одна была направлена на извлечение более ранних сообщений, а другая была предназначена для сбора более свежих записей. Собранные данные включали обсуждения политики, общественных деятелей, организаций и локаций, связанных с обозначенным кризисом. При измерении на данных с ручной разметкой, предложенный метод достиг значения полноты 0,705 и точности около 0,496 для первого набора данных, в то время как для второго набора были зафиксированы значения полноты 0,716 и точности 0,436. Основными факторами, ограничивающими точность, стали качество низлежащих моделей и ложные срабатывания: широко обсуждаемые, но непротиворечивые маркеры. Наконец, было установлено, что изучение типичного распределения эмоций в контенте социальных медиа может быть полезным для повышения качества обнаружения противоречивости.

Ключевые слова: обнаружение противоречивости, социальные медиа, обработка естественного языка, анализ тональности, распознавание именованных сущностей

Публикация подготовлена в рамках проекта РНФ № 25-28-01153 «Моделирование механизмов социального заражения в процессе онлайн-радикализации».

1. Introduction

The role of the Internet in modern interpersonal communications is as prominent as ever. Specialized social media platforms have fostered dense social networks that help disseminate vast amounts of information rapidly. The nature of information being disseminated publicly has long been of interest to socio-political researchers. One category of information that is extensively studied is the controversial content — information that invokes mixed reactions and leads to arguments, often in an uncivil manner. For example, differences in public reactions to policy and policymakers can be studied to guide future political appearances [Diakopoulos, Shamma, 2010]. Moreover, business representatives also may take interest in public opinion, as any controversies that remain unmitigated may hurt the image of the business [Amrozi et al., 2024]. Especially in the latter case it becomes evident that two controversies may not reach the same scale or produce the same impact, highlighting the need for metrics to compare and rank such events.

The analysis of social media content presents several challenges. The sheer volume of online information makes it impossible to manually review data, driving the need for automation. Another pitfall of digital media is the inherent lack of representation of actual social connections: interactions such as "follows", "friendships", "likes" or "reposts" may not reflect actual relationships between individuals. Therefore, conclusions must be drawn cautiously from the analysis of the structure of digital social media. Some additional context for relationships between individuals can be gained by analyzing the contents of social media publications, written in natural language. However, natural language processing presents its own challenges: polysemy, polymorphism and general uncertainty of human communication.

Modern machine learning tools, particularly transformer-based models, effectively model nuanced linguistic concepts. They have been successfully applied to diverse NLP tasks, from estimating emotional connotation of the text to extracting named entities that serve as conversation subjects. Thus, a significant number of models were created (or pretrained on more "general" data) and released openly to the public to be used in typical tasks.

This study aims to develop an automated controversy detection method using contemporary pre-trained language models to measure emotions expressed towards named entities. To achieve the goal, the following tasks were put forward:

1. Adapt the pre-trained models for emotion classification and named entity recognition to controversy detection.

2. Develop the emotional divergence measure to quantify controversy.

3. Conduct a case study on Reddit data to validate the proposed controversy detection pipeline and evaluate its accuracy.

2. Related work

The detection and quantification of controversial online content have been studied previously from a variety of angles. First and foremost, most definitions of controversy, controversiality or controversial topic given by many authors include some characteristic of dispute or difference in opinion. To define the controversy more precisely, the related phenomenon of polarization must be reviewed. One particular work [Bramson et al., 2016] describes polarization in 9 "senses". Of those, the "senses" of spread of opinion and dispersion of attitudes fit earlier notion of controversy. Examining further works on the phenomenon in discussions, it is revealed that the polarization is characterized as a split into opposing groups [Al Amin et al., 2017; Guerra et al., 2021], in line with the description of controversy. The difference between the terms seems to stem from the goals of the researchers:

polarization is used by authors who seek to identify such opposing groups in a social graph, while controversy is used by authors who attempt to classify subjects into controversial or not. In the end, this work has defined the controversial content as any content that invokes mixed responses, and therefore, disagreements. In an extreme case, the disagreement may devolve into uncivil exchanges that are explicitly included into the present definition. For brevity, the characteristic of content to have such a disagreement in attitudes is referred to as controversy.

Researchers have set different goals when tackling the phenomenon of controversy. Some works aim to classify a particular known topic as controversial or not [Mejova et al., 2014; Dori-Hacohen, Allan, 2015; Jang et al., 2016]. On the other hand, several researchers seek to identify all instances of controversy in the dataset of user communications without significant prior knowledge [Gomez et al., 2008; Popescu, Pennacchiotti, 2010; Coletto et al., 2017]. This work fits into the latter category and emphasizes automation in identification and quantification of all instances of controversy.

Two categories of approaches to the task are identifiable: structural and semantic. Structural approaches tend to analyze the social network — a graph of users and user generated content. As such, entire networks may be classified as controversial based on global graph properties [Guerra et al., 2021; Benslimane et al., 2023]. On the contrary, some works study local graph features to identify particularly controversial parts [Gomez et al., 2008; Coletto et al., 2017]. Structural methods are not tied to the language of the discussion, bypassing the difficulty of comprehending the exact content of the discussion. Yet any implementation of a structural method necessarily uses additional information for building social networks, such as user relations to build edges.

On the contrary, semantic methods delve into the actual content being discussed. Some semantic approaches also use metadata like social media hashtags to relate messages or record endorsement [Garimella et al., 2018; Hellsten, Leydesdorff, 2020]. Wikipedia-based controversy detection deserves special attention as it seeks to relate web content to Wikipedia articles based on the most frequent terms [Dori-Hacohen, Allan, 2015; Jang et al., 2016]. Then, the controversy of the text can be related to known controversy metrics of Wikipedia articles. Naturally, the main obstacle becomes the availability of Wikipedia material and its variety.

More relevant to this study are the attempts to gauge controversy via sentiment analysis, which estimates positive and negative attitudes towards content. Researchers have tried to measure and classify controversial messages by such scalar sentiment values that are usually bound between -1 and 1. Older approaches utilized specific dictionaries that assign a sentiment value to words [Popescu, Pennacchiotti, 2010; Goncalves et al., 2013; Mejova et al., 2014]. Using such dictionaries, messages were processed word by word and had their aggregated sentiment calculated based on tokens' scores. More modern approaches produce sentiments by leveraging natural language processing tools: conditional random field models and large language models (mainly encoders) [Qiu et al., 2019; Benslimane et al., 2023].

After acquiring sentiment values for content, some researchers utilize them to build or refine social networks, creating a mixed structural-semantic approach to controversy detection [Garimella et al., 2018; Qiu et al., 2019; Benslimane et al., 2023]. Other works explore relation between sentiment and controversy directly. As such, several researchers have concluded that there is a correlation between negative sentiment and controversy [Lheureux, 2024; Qian et al., 2025]. One study [Diakopoulos, Shamma, 2010] has found that both negative and positive sentiment correlate to controversy in context of political debates. Others have studied the distribution of sentiment as a measure of controversy, more in line with the definition of controversy as difference of opinion. Specifically, dispersion thresholds were successfully used to identify controversy [Tsytsarau et al., 2011; Garimella et al., 2018].

Semantic methods have a significant drawback: dependence on language. The need to create dictionaries for every processed language has limited the usefulness of the approach in the past. However, modern large language models can be trained to recover sentiments on multilingual datasets,

bypassing the language barrier. Still, one study [Dori-Hacohen, Allan, 2015] has concluded that sentiments are rarely enough to identify controversy, although others have expanded upon this statement and have shown that the sentiment distribution serves as a better evidence of controversy [Garimella et al., 2018].

A promising advancement in sentiment analysis is the advent of emotion classification models. Such modern tools of natural language processing seek to embed the input text into "emotional space" — a vector with components that encode the degree to which a particular emotion is expressed. Unlike earlier sentiment models that produce a scalar value for text, emotional classification models may uncover more subtle variations in content, possibly improving the detection of controversy based on dispersion of sentiment. To the best of authors' knowledge, such models are yet to be applied for the task of controversy detection. Although the emotional aspect of online interaction is being studied [Qian et al., 2025].

Moreover, tools of named entity recognition have not been used extensively to identify controversial topics. Past works either relied on prior knowledge of data to extract subjects of discussion or used simpler algorithms, such as latent Dirichlet allocation [Tsytsarau et al., 2011].

In the end, this work seeks to explore modern tools of natural language processing, such as named entity recognition and emotion classification models, and expand upon earlier semantic methods of controversy identification.

3. Controversy detection pipeline

Emotion-based approach for controversy detection is designed to operate on social media data. Any chosen social media dataset must include records of messages, each with identifiers and text content. The dataset D with N datapoints that represent messages is defined as:

$$D = \{x_i \mid i = 1, 2, \ldots, N\}. \tag{1}$$

The data acquisition is the first step towards controversy detection. Then, the data can be processed as described further in Fig. 1.

The pipeline first employs sentiment analysis models, which are machine learning tools that estimate the emotional connotation of text. Given message text x as input, these models produce emotion vectors $\boldsymbol{e}$ in emotion space $\mathbb{R}^m$. The dimensionality m of the emotion space is fixed and depends on the model's architecture. Importantly, the output vector is normalized, meaning that the sum of its components is always 1. As such, the component's value can be interpreted as the proportion of the input message that conveys the given emotion. The transformation f from text x to emotion space $\mathbb{R}^m$ can be described by the equation

$$\boldsymbol{e} = f(x), \quad x \in D, \quad \boldsymbol{e} \in \mathbb{R}^m, \quad \sum_{j=1}^{m} e_j = 1. \tag{2}$$

In addition to sentiment analysis, named entity recognition is performed on the messages in the dataset using specialized machine learning models. Named entities are usually proper nouns that often take the place of the subject or the object of a sentence. Thus, named entities serve as an estimate of the topics that were discussed in conversations. Models in question produce a list $\boldsymbol{n}$ of lexemes that are deemed named entities along with the category and confidence of identification for each list item. It must be noted that the resulting list can be arbitrarily long. The confidence score of each list item l is a real value that is bound to the range [0, 1]. Let $c(l)$ be a function that outputs the confidence score of the named entity list item. Then, the named entity recognition transformation g from text x to named entity list $\boldsymbol{n}$ can be described by the equation

$$\boldsymbol{n} = g(x), \quad \forall l \in \boldsymbol{n},\ c(l) \in [0, 1]. \tag{3}$$

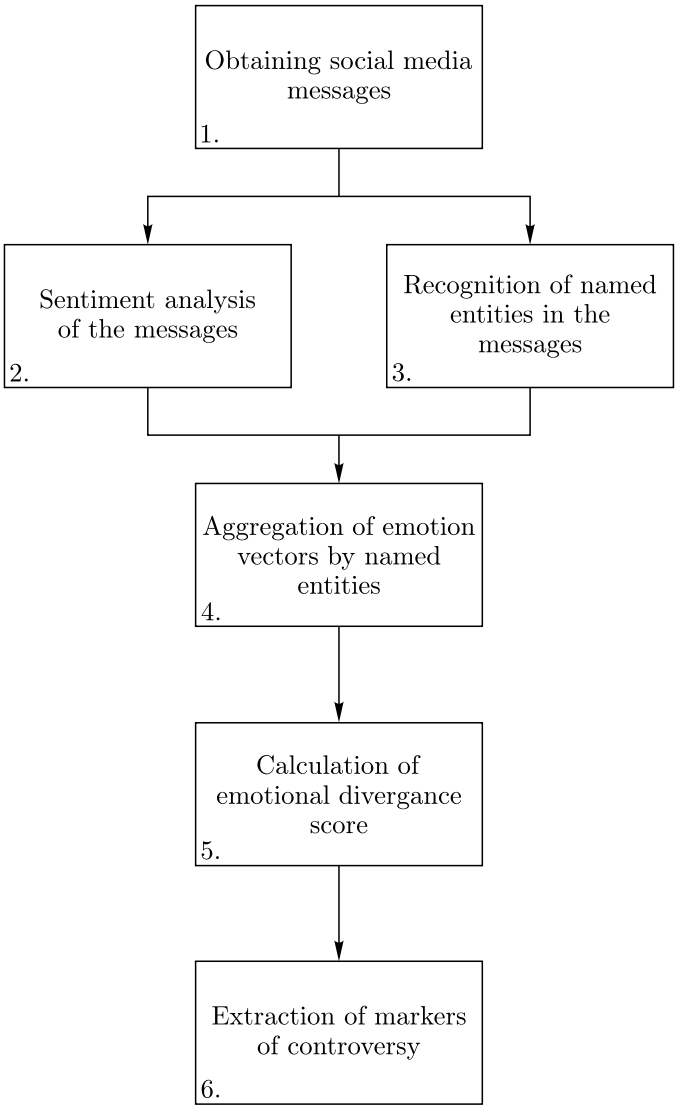

Figure 1. A diagram of the pipeline

Named entity recognition models tend to extract semantically insignificant lexemes. To improve the accuracy of analysis, these outliers must be discarded. Entities with low confidence scores tend to be wrongly identified. As such, it was decided to exclude any lexeme with a confidence score lower than the dataset-wide median from the analysis. The median threshold was chosen to reduce the number of identified named entities by half. The retrieved entities should also be manually reviewed to remove semantically insignificant lexemes, such as misidentified prepositions and pronouns. A degree of expertise in the subject area is also helpful to adequately filter the entities.

Once the emotion vectors and named entities were calculated for all samples x_i in the dataset D, the dataset-wide statistics of emotional connotations for each entity were explored. Firstly, an aggregation mapping t that relates named entity l from dataset-wide set of unique named entities L to a list $\boldsymbol{E}$ of associated emotion vectors $\boldsymbol{e}$ was constructed.

Definition 1. Named entity l is considered associated with emotion vector $\boldsymbol{e}$ if and only if there exists a sample x_i in the dataset D such that $e = f(x_i)$, $l \in g(x_i)$.

Consequently, several samples may contribute emotion vectors to the aggregate of one entity, in case such samples mention the named entity in question.

Then, for every mapped named entity with at least 2 emotion vectors a scalar value of emotional divergence was calculated. Emotional divergence score aims to estimate how dissimilar the emotional connotations of messages about any given entity are.

Definition 2. Emotional divergence $d(l)$ of a named entity l is defined as the sum of standard deviations of all m vector components in the list of associated emotion vectors $t(l)$. Since $\boldsymbol{E} = t(l)$ is a matrix (as a list of n row vectors $\boldsymbol{e}$ in $\mathbb{R}^m$), the emotional divergence can be expressed as

$$\boldsymbol{E} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}, \quad d(l) = \sum_{j=1}^{m} \sigma(\{e_{1j}, e_{2j}, \ldots, e_{nj}\}) \tag{4}$$

where σ represents the standard deviation of the column $\{e_{1j}, e_{2j}, \ldots, e_{nj}\}$. The standard deviation was chosen as a nonnegative variance measure. It is notable that adding an associated vector to any entity will produce a nonnegative term in the sum, naturally increasing the emotional divergence score of named entities with a lot of discussion. This effect is desirable, since discussions with many participants tend to disagree more than limited conversations.

Standard deviation is not the only statistical measure of the distribution that can capture polarized opinions. For example, the interquartile range can be used as a measure of variance instead. Yet interquartile range is robust to outliers, making it less desirable for controversy detection, as extreme emotional responses often serve as key indicators. Moreover, the computation of interquartile range is also more expensive than that of standard deviation due to the requirement of sorting the data. That makes standard deviation more suited to processing of large datasets.

Furthermore, the measures of informational uncertainty of distributions may be considered as a substitute for dispersion measures in emotional divergence computation. For example, it is possible to define emotional divergence as the sum of entropies of all emotions' distributions. The main concern with using entropy is methodological: the measure of uncertainty can only capture differing opinions indirectly, as less determined distributions. Moreover, there is a concern with numerical stability of entropy computation, requiring some preprocessing of distribution data.

All in all, the standard deviation not only fits controversy detection methodology by directly capturing dispersed and polarized attitudes, but is also more computationally efficient.

Lastly, the distribution of emotional divergence scores is studied to identify entities that are mentioned in messages with dissimilar emotional connotations. Such entities are designated as markers of controversy, as they are used to predict which conversations are likely to become controversial. The specific threshold of emotional divergence score that identifies markers of controversy is derived empirically, usually taking on a value above the median emotional divergence. Additionally, list of markers can be manually reviewed to exclude entities that are too broad to be deemed controversial.

4. Experiments

An experiment was devised to verify the ability of the emotion-based pipeline to identify controversial content in social media. The social media platform known as Reddit was chosen as the data source due to the public availability of its API and primarily English communications. Specifically, data extraction efforts were focused on the English-speaking community r/srilanka that is dedicated to the discussions of life in the country of Sri Lanka. The community was chosen due to the belief that it may have harbored controversial sentiments from the 2022 political crisis in the country.

Data extraction pipeline uses Reddit search mechanism to identify submissions marked by requested flairs or containing requested keywords. Submissions serve the role of root messages within

the Reddit data model, setting the initial topic of discussion. Once submissions of interest are identified, the replying comments are recursively extracted to expand the dataset.

Some additional considerations must be made about the search parameters. Flairs that categorize submissions are defined on per-community basis by moderators, so some expertise is required to select meaningful flairs. Moreover, the community r/srilanka has only adopted the widespread use of flairs to mark submissions in June of 2023, meaning that earlier messages cannot be extracted by flairs. To counteract this issue, an additional way of searching for submissions was considered by occurrence of keywords. However, the choice of specific keywords still requires expertise in the subject area.

The resulting dataset aggregates both submissions and comments into linear records that represent tree-like structures of replies. Several additional fields are included to make dataset's origin verifiable. The scheme of the dataset is described in Table 1.

Table 1. Dataset scheme

Field name	Description	Nullable	Type
full_name	Reddit API ID of the record.	No	String
text_body	Text content of the record, encoded in UTF-8.	Yes	String
author_name	Reddit API ID of the record's author.	Yes	String
votes	User-defined rating of the record.	No	Integer
responds_to	Reddit API ID of the record being responded to.	Yes	String
parent_submission_name	Reddit API ID of the record's submission.	Yes	String
submission_flair	Submission flair of the record.	Yes	String
created_timestamp	UNIX timestamp of record creation.	No	Integer
parsed_timestamp	UNIX timestamp of record parsing.	No	Integer
controversiality	Reddit API flag of controversiality.	No	Boolean

A Python application was developed with use of PRAW package. The software expects a list of flairs and/or keywords to search the community by. If flairs are provided, submissions with the given flairs are extracted. If keywords are provided, submissions that contain the keywords and lack any flairs are extracted. This feature serves to accommodate the search of pre-2023 submissions that do not contain any flairs. The extracted data is stored in a CSV file.

The first dataset was collected with 5 flairs: "Politics", "News", "Discussion", "Bureaucracy" and "Rumor". These flairs were chosen in an effort to capture the greatest number of controversial discussions. In total, 51 892 records, authored by 7512 accounts among 922 submissions, were collected in 30 minutes. In addition to submissions with the flair "Rumor", submissions with the flair "Rumor Disproven" were also present in the resulting dataset. Submissions were not equally distributed among flairs: "News" had 243 submissions, "Politics" — 238, "Discussion" — 235, "Rumor" — 189, "Bureaucracy" — 11, "Rumor Disproven" — 5. The Reddit's controversiality flag was set only for 2936 records, marking 5 % of the data. It is believed that this flag underestimates the number of controversial messages. User votes were distributed unequally, as expected of social media data. While the maximum vote rating was above 90, more than 75 % of records had 6 votes or fewer. Furthermore, some correlation between votes and the Reddit controversiality flag is identifiable, as controversial records generally have the rating of 1 or less. Lastly, 85 % of all records were published in the year of 2024, highlighting the effect of the adoption of flairs by the community.

Then, the flair dataset was studied for markers of controversy. The open model "Babelscape/wikineural-multilingual-ner" [Tedeschi et al., 2021] was chosen for the task of named entity recognition, mainly due to the capability of multilingual analysis. The model identified 74 757 entities across the records of the dataset. It was found that 17 275 unique named entities occur in the data. The study of confidence scores of model's predictions revealed high certainty of the model,

as was evident by the median score of 0.8227. Thereafter, the list of considered entities was reduced to 8226 unique instances that have a higher confidence score than the median.

In parallel, the emotion vectors were computed for the records in the dataset with the help of the open model "j-hartmann/emotion-english-distilroberta-base" [Hartmann, 2022]. The model features 7 emotion components that include 6 basic emotions by Paul Ekman and a neutral sentiment. Moreover, the model was partly trained on data from Reddit, substantiating its choice for the dataset at hand. Examination of produced emotion vectors revealed the dominance of neutral component, as the median value for it exceeded 0.50. All other emotions had median values of approximately 0.04 or below. Still, the dispersion of distributions for each vector component was found to be significant, as shown by the maximum values of 0.99 and higher. The distributions were visualized in the form of boxplots, as shown in Fig. 2.

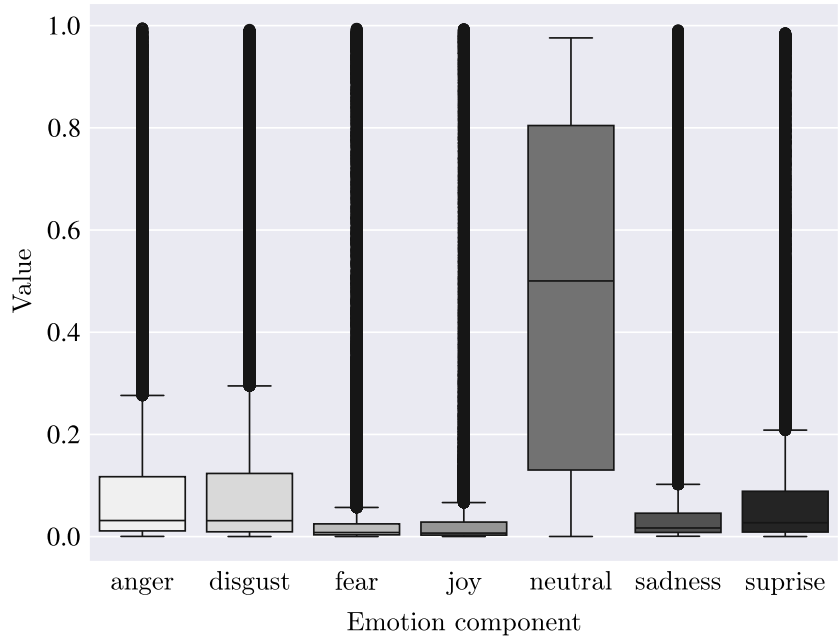

Figure 2. Emotion component distributions of flair dataset

As emotion vectors were aggregated by associated named entities, 2261 entities with 2 or more emotions were discovered. The emotional divergence of such entities was found to be distributed between 0.00 and 1.61. The data was further split into two nonintersecting subsets for visualization and testing at an 80 : 20 ratio, for a more objective evaluation of the pipeline. The visualization split was named like that because no training parameters are fitted with it, making the classic machine learning term "training split" ill-fitted.

While exploring the visualization data split, 25 % of the entities were found to have an emotional divergence score of 0.28 and lower, marking them as unlikely controversial. Furthermore, entities were more densely distributed around the emotional divergence value of 0.9, while being sparsely distributed further out. The shape of the distribution of emotional divergence scores was plotted in the form of a histogram as presented in Fig. 3.

Based on the visualization data split, a value of 0.60 was chosen as a threshold for controversy, meaning that only the entities with the emotional divergence score above 0.60 were selected as markers of controversy. The value of threshold was chosen in an effort to capture the more disputable entities that are distributed in the "peak" area of the plot, while leaving out the less controversial entities in the "plateau" area to the left of the plot.

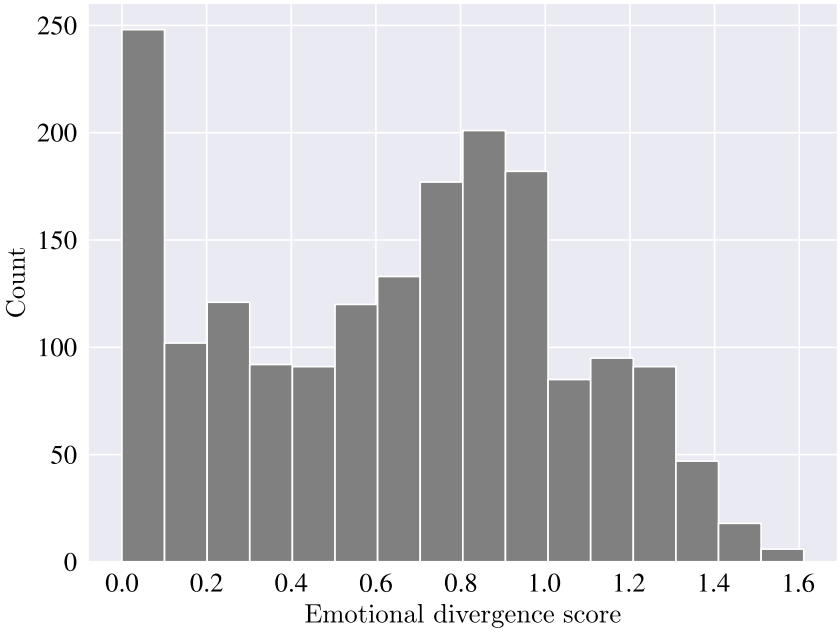

Figure 3. Emotional divergence distribution for visualization split of flair dataset

Then, the classification accuracy of the pipeline with the chosen threshold was measured against the ground truth. For every entity in a testing data split, a list of messages that mention the entity was compiled. Each list was manually reviewed and labeled as either controversial or not. The entity was deemed controversial if at least one of two things was evident: messages exhibited a difference in opinion or messages contained uncivil arguments (mainly ad hominem expressions with little relation to the subject). Additionally, words with little lexical meaning (like auxiliary verbs) were labeled as not controversial.

The testing split contained 452 entities in total, with 176 controversial labels and 276 with uncontroversial labels. Classification based on the proposed threshold of 0.6 has yielded the results as shown in Table 2.

Table 2. Flair dataset classification on testing split with 0.6 threshold

	Labeled controversial	Labeled non-controversial
Predicted controversial	124	126
Predicted noncontroversial	52	150

In summary, the precision of 0.496 and recall of 0.705 was achieved with a threshold of 0.6. The Performance with various thresholds was studied and was visualized with a ROC curve. The AUC was measured at 0.698. The curve is shown in Fig. 4.

Then the threshold of 0.6 was used to classify the entirety of 2261 entities, producing 1288 markers of controversy. After interjections and common phrases were manually removed, a list of 1240 markers was finalized. 200 entities with the highest emotional divergence scores were plotted as a word cloud for demonstration of the pipeline's output, as shown in Fig. 5.

Another experiment was conducted with new data to verify findings. The keyword dataset was collected with the aim to capture more controversial discussions from the 2022 political crisis. For this dataset, submissions were extracted by 16 keywords that relate to locations, individuals, or concepts that were significant in the timeframe. The keywords consisted of: "2022", "colombo", "cost", "crisis", "economic", "galle", "gota", "gotabaya", "kohuwala", "martial", "mirihana", "protest",

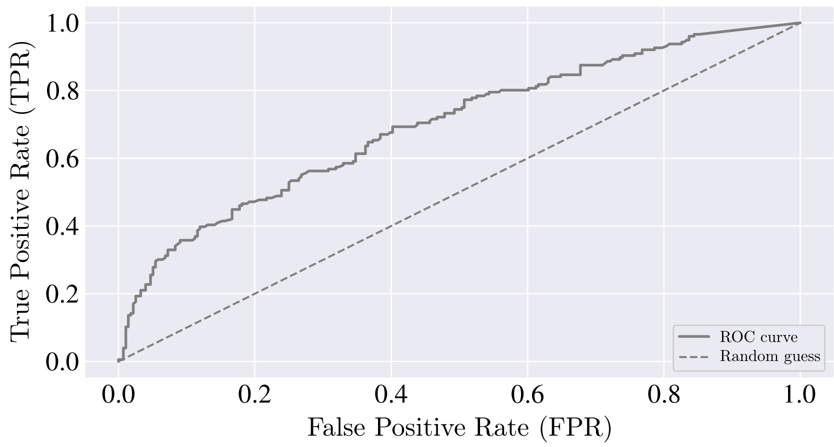

Figure 4. ROC curve for classification on testing split of flair dataset

Figure 5. Word cloud of controversial entities for flair dataset

“rajapaksa”, “riot”, “sjb” and “storm”. This time, 24 606 records, authored by 4390 accounts among 1569 submissions, were collected in 18 minutes. Similar to the first dataset, only around 3 % of records were marked with the Reddit controversiality flag in the second dataset. User votes were also distributed with a skew, with 75 % of records having 5 or fewer votes. However, 50 % of all records did belong to the year 2022, as desired.

Subsequently, the keyword dataset was also studied for markers of controversy. The same open model “Babelscape/wikineural-multilingual-ner” [Tedeschi et al., 2021] was used for named entity recognition, ensuring the comparability of the results with the previous experiment. The model identified 36 849 entities, belonging to a set of 10 588 unique entries, across the records of the dataset. The study of confidence scores of model’s predictions supported earlier findings of model’s high certainty, as was evident by the median score of 0.8565. Afterwards, the number of named entities was reduced to 4732 unique entries with respect to the median as the threshold.

Likewise, the same open model “j-hartmann/emotion-english-distilroberta-base” [Hartmann, 2022] was selected for sentiment analysis. Review of emotion vectors once again showed dominance of neutral sentiment, with the median value in excess of 0.54. All other emotions had median values of

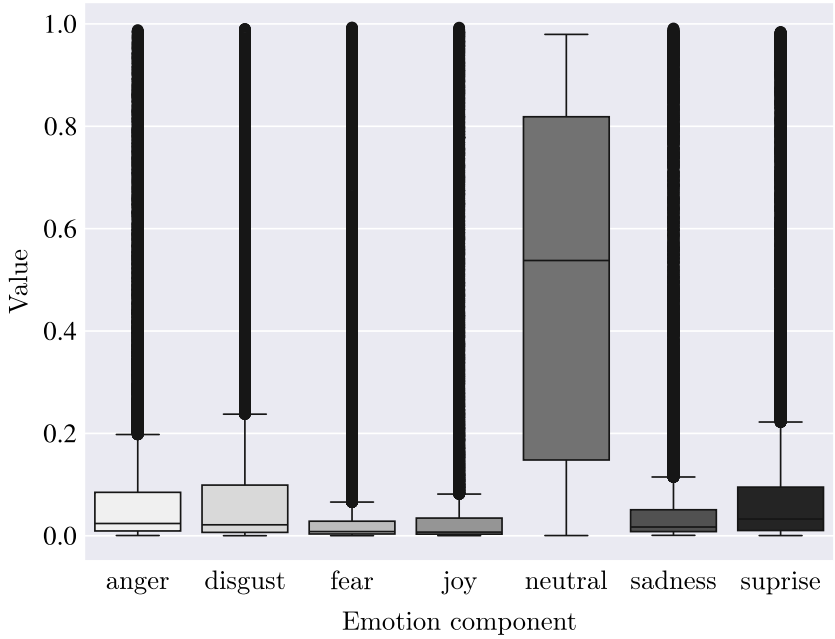

Figure 6. Emotion component distributions of keyword dataset

approximately 0.03 or below. Furthermore, maximum values of 0.98 and higher corroborate significant dispersions of distributions that were observed earlier. The emotion distributions of keyword dataset were visualized in the form of boxplots, as shown in Fig. 6.

Afterwards, emotion vectors were aggregated, resulting in 1277 entities with 2 or more associated vectors. The emotional divergence of such entities was found to be distributed between 0.00 and 1.59 in a similar fashion to the previous experiment. The data was split into visualization and testing sets at an 80 : 20 ratio.

Studying the visualization split, the noncontroversial "peak" in the first quarter, the "plateau" region and the controversial "peak" with dense distribution were all observed. The shape of the distribution of emotional divergence scores of keyword dataset was plotted in the form of a histogram as illustrated in Fig. 7.

A value of 0.6 was chosen as the threshold of controversy based on the same reasoning as in the flair experiment, allowing for validation of the threshold's performance on different data. Moving on to accuracy evaluation, the experiment on keyword dataset, in turn, uncovered 1277 emotionally divergent entities. As such, the 20 % split contained 255 entities in total, with 81 controversial labels and 174 with uncontroversial labels. Classification based on the threshold of 0.6 has yielded the results as shown in Table 3.

Table 3. Keyword dataset classification on testing split with 0.6 threshold

	Labeled controversial	Labeled non-controversial
Predicted controversial	58	75
Predicted non-controversial	23	99

Thus, the precision of 0.436 and recall of 0.716 was achieved on testing split. The summary of performance was expressed as a ROC curve with 0.723 AUC, as shown in Fig. 8.

Next, the 0.6 threshold was applied to the entire keyword dataset. As such, 726 markers of controversy were produced for a manual review. After common phrases were removed, a list

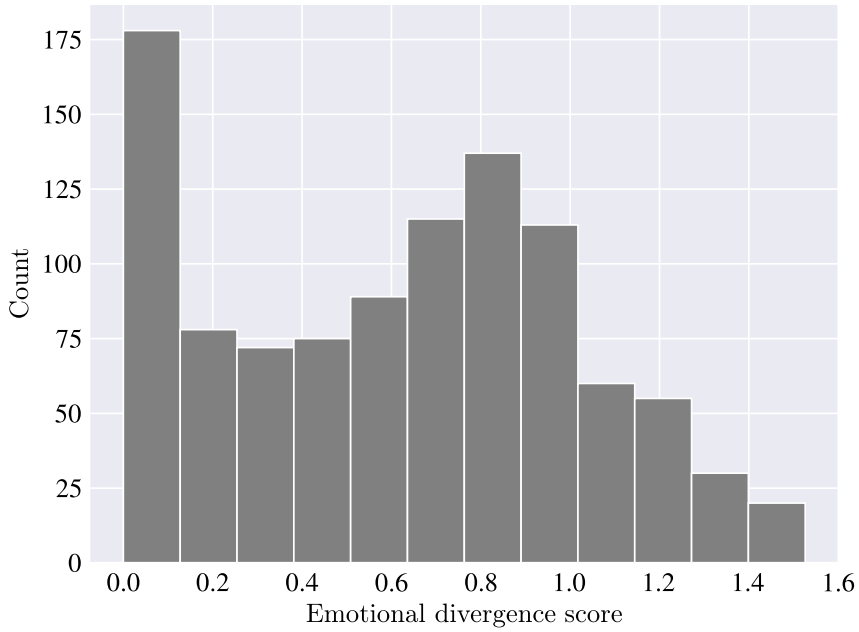

Figure 7. Emotional divergence distribution for visualization split of keyword dataset

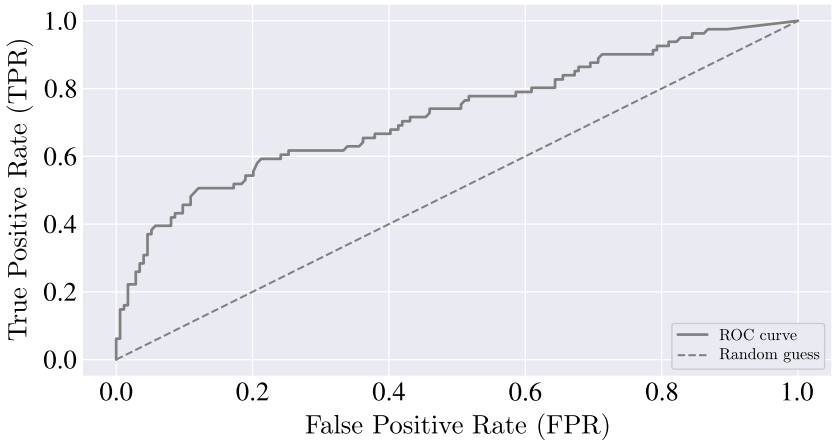

Figure 8. ROC curve for classification on testing split of keyword dataset

of 713 markers was finalized. Lastly, 200 entities with the highest emotional divergence scores among the keyword dataset were plotted as a word cloud for demonstration, as shown in Fig. 9.

5. Discussion

In both case studies it was found that the proposed method achieves sufficient recall, while the precision remains relatively low. This can be explained by a significant false positive rate: evidently not only the controversial entities receive mixed emotional response. Looking closely at the false positives, there are discussions about things of personal preference (cars, books, tourist attractions and others) where users express largely the same sentiment with differing emotional language.

On the other hand, high recall may be desirable in the context of social media analysis as uncovering the full variety of controversial markers (that is, topics of discussion) helps construct a more complete image of discussions. On the other hand, the achieved level of precision necessitates

Figure 9. Word cloud of controversial entities for keyword dataset

effort from subject area experts to exclude false positives. Thus, the improvement to precision remains a challenge.

Direct comparison to previous semantic controversy detection methods is also challenging. Several approaches assume specific social media structure and cannot be easily applied to the collected data of Reddit case studies. Moreover, the varying definitions of controversy across studies further complicate meaningful comparisons, as each approach operates under different assumptions. When comparing the most relevant works to the emotional divergence pipeline, basically every other approach exceeds the original method as per reported metrics. The comparison is shown in Table 4.

Table 4. Controversy detection comparisson

	Accuracy	Precision	Recall	AUC
Benslimane et al., 2023	0.638–0.748			
Coletto et al., 2017		0.770–0.870	0.820–0.880	
Jang et al., 2016				0.795–0.801
Popescu, Pennacchiotti, 2010				0.805–0.850
Original pipeline	0.606–0.616	0.436–0.496	0.705–0.716	0.698–0.723

A closer study of controversy markers and their discussions reveals a pattern: controversial entities tend to have more associated messages, which was accounted for in the method's design. Intuitively, the more responses the topic gets the more opportunities there are for an opposing opinion to emerge, satisfying the controversy criteria. Yet, such a correlation has very evident counterexamples: topics that devolve into uncivil arguments within just 2 messages (especially prevalent in political discussions). In general, nearly every entity related to politics that was manually tested was found to be controversial.

Moving on to limitations, the pretrained language models used for emotion classification and named entity recognition have limited accuracy. Therefore, the proposed method of controversy detection can only be as accurate as all underlying models. Moreover, all language models are trained on some specific data, that can introduce biases to the estimations made by any particular model. As such, special attention must be paid to the selection of the model: better performance may be achieved if the model is used on texts that are similar to model's training data. The languages that models can

process also depend on the training data. Although many models are trained on multilingual datasets, the performance of such models may still vary based on the language of the text.

More specifically, the "Babelscape/wikineural-multilingual-ner" model has F1 measure between 0.6 and 0.9 as reported in the study [Tedeschi et al., 2021]. It must be noted that lower values were recorded for more free form texts that deviate from Wikipedia's formality. Similarly, the "j-hartmann/emotion-english-distilroberta-base" was estimated to have 0.66 accuracy (rate of correctly predicted labels) in the work [Hartmann, 2022].

Furthermore, although diverging emotions can uncover differences in opinion and thus controversy, some controversial discussions may occur entirely within a single emotional category (for example, "anger"). Especially the least constructive conflicts may devolve into mutual threats that are likely to be embedded closely in the "emotional space". Emotional divergence is not suited to identify such cases of controversy.

Although the proposed method automates controversy detection, some human intervention remains necessary. Produced markers of controversial discourse should be reviewed by subject area experts to evaluate their usefulness. Moreover, possibly only such experts can conclude whether or not the markers have the generalizing ability (that is, could the markers produced from one dataset be used to detect controversy in other data).

6. Conclusion

In the end, a controversy detection and quantification pipeline based on named entities and emotions was proposed. Its effectiveness was measured in two case studies on Reddit data and was deemed sufficient for detection, as the number of markers that must be manually reviewed was reduced to manageable levels (thousands of markers). Several tendencies in the produced markers of controversy were identified: controversy increases with the number of messages, political entities are controversial and there are emotional false positives entities with no controversy.

It is expected that improvements to pretrained models could enhance the overall accuracy. Additionally, the estimation of base emotional divergence across large portions of social media (the figurative "emotional background") seems to be a prospective way to eliminate false positives. By accounting for entities that are emotionally divergent in the same way as the typical content, controversies may become more evident as anomalies against the "emotional background".

References

Al Amin M. T., Aggarwal C., Yao S., Abdelzaher T., Kaplan L. Unveiling polarization in social networks: A matrix factorization approach // IEEE INFOCOM 2017 – IEEE Conference on Computer Communications. — IEEE, 2017. — P. 1–9. — https://doi.org/10.1109/infocom.2017.8056959

Amrozi A. I., Ghofur A., Damayanti D. D., Suprapto H., Sulaeman M. M. Dynamics of social media interaction: Implications for the manifestation of digital business image and reputation in public perception // Tacit. — 2024. — Vol. 2, No. 1. — P. 150–157. — https://doi.org/10.61100/tacit.v2i1.141

Benslimane S., Azé J., Bringay S., Servajean M., Mollevi C. A text and GNN based controversy detection method on social media // World Wide Web. — 2023. — Vol. 26, No. 2. — P. 799–825. — https://doi.org/10.1007/s11280-022-01116-0

Bramson A., Grim P., Singer D. J., Fisher S., Berger W., Sack G., Flocken C. Disambiguation of social polarization concepts and measures // The Journal of Mathematical Sociology. — 2016. — Vol. 40, No. 2. — P. 80–111. — https://doi.org/10.1080/0022250X.2016.1147443

Coletto M., Garimella K., Gionis A., Lucchese C. A motif-based approach for identifying controversy // Proceedings of the International AAAI Conference on Web and Social Media. — 2017. — Vol. 11, No. 1. — P. 496–499. — https://doi.org/10.1609/icwsm.v11i1.14949

Diakopoulos N. A., Shamma D. A. Characterizing debate performance via aggregated twitter sentiment // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. — New York, NY, USA: ACM, 2010. — P. 1195–1198. — https://doi.org/10.1145/1753326.1753504

Dori-Hacohen S., Allan J. Automated controversy detection on the web // Lecture Notes in Computer Science. — Cham: Springer International Publishing, 2015. — P. 423–434. — https://doi.org/10.1007/978-3-319-16354-3_46

Garimella K., Morales G. D. F., Gionis A., Mathioudakis M. Quantifying controversy on social media // ACM Transactions on Social Computing. — 2018. — Vol. 1, No. 1. — P. 1–27. — https://doi.org/10.1145/3140565

Gomez V., Kaltenbrunner A., Lopez V. Statistical analysis of the social network and discussion threads in slashdot // Proceedings of the 17th International Conference on World Wide Web. — New York, NY, USA: ACM, 2008. — P. 645–654. — https://doi.org/10.1145/1367497.1367585

Gonçalves P., Araújo M., Benevenuto F., Cha M. Comparing and combining sentiment analysis methods // Proceedings of the First ACM Conference on Online Social Networks. — New York, NY, USA: ACM, 2013. — P. 27–38. — https://doi.org/10.1145/2512938.2512951

Guerra P., Meira Jr. W., Cardie C., Kleinberg R. A measure of polarization on social media networks based on community boundaries // Proceedings of the International AAAI Conference on Web and Social Media. — 2021. — Vol. 7, No. 1. — P. 215–224. — https://doi.org/10.1609/icwsm.v7i1.14421

J-hartmann/emotion-english-distilroberta-base. — [Electronic resource]. — https://huggingface.co/j-hartmann/emotion-english-distilroberta-base (accessed: 21.10.2025).

Hellsten I., Leydesdorff L. Automated analysis of actor-topic networks on twitter: New approaches to the analysis of socio-semantic networks // Journal of the Association for Information Science and Technology. — 2020. — Vol. 71, No. 1. — P. 3–15. — https://doi.org/10.1002/asi.24207

Jang M., Foley J., Dori-Hacohen S., Allan J. Probabilistic approaches to controversy detection // Proceedings of the 25th ACM International Conference on Information and Knowledge Management. — New York, NY, USA: ACM, 2016. — P. 2069–2072. — https://doi.org/10.1145/2983323.2983911

Lheureux Y. Predictive insights: leveraging Twitter sentiments and machine learning for environmental, social and governance controversy prediction // Journal of Computational Social Science. — 2024. — Vol. 7, No. 1. — P. 23–44. — https://doi.org/10.1007/s42001-023-00228-5

Mejova Y., Zhang A. X., Diakopoulos N., Castillo C. Controversy and sentiment in online news // arXiv preprint. — 2014. — https://doi.org/10.48550/ARXIV.1409.8152

Popescu A.-M., Pennacchiotti M. Detecting controversial events from twitter // Proceedings of the 19th ACM International Conference on Information and Knowledge Management. — New York, NY, USA: ACM, 2010. — P. 1873–1876. — https://doi.org/10.1145/1871437.1871751

Qian Y., Zhao J., Wang G. Adaptive simulation analysis of users' emotional information dissemination network in the English context of integrated media platforms // International Journal of Housing and Its Applications. — 2025. — Vol. 46, No. 4. — P. 3108–3120. — https://doi.org/10.70517/ijhsa464257

Qiu J., Lin Z., Shuai Q. Investigating the opinions distribution in the controversy on social media // Information Sciences. — 2019. — Vol. 489. — P. 274–288. — https://doi.org/10.1016/j.ins.2019.03.041

Tedeschi S., Maiorca V., Campolungo N., Cecconi F., Navigli R. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER // Findings of the Association

for Computational Linguistics: EMNLP 2021. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2021. — https://doi.org/10.18653/v1/2021.findings-emnlp.215

Tsytsarau M., Palpanas T., Denecke K. Scalable detection of sentiment-based contradictions // DiversiWeb 2011 Proceedings of the 1st International Workshop on Knowledge Diversity on the Web Workshop at the 20th International World Wide Web Conference WWW 2011. — 2011. — P. 9–16.