OPINION WG2 - Coding error analysis

Johannes B. Gruber

2023-10-23

Table of contents

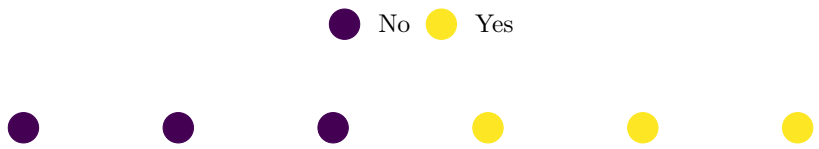
1 Relevant 1 **1**

2 Concept **8**

3 relevant-2 **28**

This document contains abstracts where agreement on whether they are relevant was especially low. We use this to see what is still unclear and where our codebook needs improvement.

1 Relevant 1

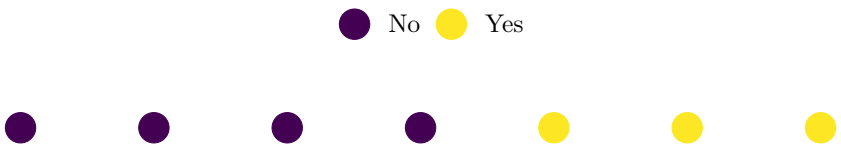


ID: 2378

Agreement Score: 0

Towards a Service-Oriented Architecture for Pre-processing Crowd-Sourced Sentiment from Twitter

Online social media platforms like Twitter, provide **opinion** rich repositories for conducting sentiment analysis. Users engage in open discussions on a variety of topics across a wide cross-section of problem domains. Commercial, government, educational, non-profit and other types of agencies are increasingly relying on extracting conversations on Twitter to determine the general sentiment of the public on particular topics, products, services and issues. Despite being readily available and in abundance, it is also laced with nuances which can disrupt, skew and potentially lead to inaccurate analysis if not handled properly. In this paper, we propose an SOA framework to enable the pre-processing of data origination on Twitter, and configurable components that allow data consumers to filter the data using useful social media signals.

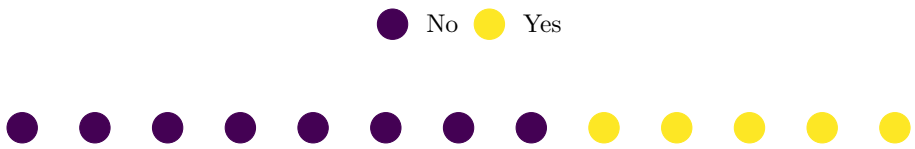


ID: 1799

Agreement Score: 0.14

Review article: Detection of actionable tweets in crisis events

Messages on social media can be an important source of information during crisis situations. They can frequently provide details about developments much faster than traditional sources (e.g., official news) and can offer personal perspectives on events, such as **opinions** or specific needs. In the future, these messages can also serve to assess disaster risks. One challenge for utilizing social media in crisis situations is the reliable detection of relevant messages in a flood of data. Researchers have started to look into this problem in recent years, beginning with crowdsourced methods. Lately, approaches have shifted towards an automatic analysis of messages. A major stumbling block here is the question of exactly what messages are considered relevant or informative, as this is dependent on the specific usage scenario and the role of the user in this scenario. In this review article, we present methods for the automatic detection of crisis-related messages (tweets) on Twitter. We start by showing the varying definitions of importance and relevance relating to disasters, leading into the concept of use case-dependent actionability that has recently become more popular and is the focal point of the review paper. This is followed by an overview of existing crisis-related social media data sets for evaluation and training purposes. We then compare approaches for solving the detection problem based (1) on filtering by characteristics like keywords and location, (2) on crowdsourcing, and (3) on machine learning technique. We analyze their suitability and limitations of the approaches with regards to actionability. We then point out particular challenges, such as the linguistic issues concerning social media data. Finally, we suggest future avenues of research and show connections to related tasks, such as the subsequent semantic classification of tweets.

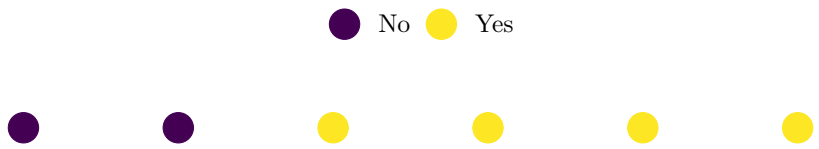


ID: 1533

Agreement Score: 0.23

Media portrayal of hackers in China Daily and The New York Times: A corpus-based critical discourse analysis

This study draws on a synergy of Corpus Linguistics and Critical Discourse Studies to scrutinize the portrayal of hackers in China Daily and The New York Times in the 21st century (2001-2020), primarily revolving around the main social actors and targets in hacking. This study demonstrates that both media share a positive transformation of the image-building of hackers in the 21st century. Besides, countries are salient social actors in hacker media discourse and the two media differ in their ways of constructing them. The New York Times tends to have a negative other-representation and categorical otherness of specific countries through such discursive strategies as negative other-representation and group categorization, whereas China Daily is prone to insist on opposing the US hacking allegations in a defensive manner. Regarding major targets, China Daily highlights government websites whereas The New York Times emphasizes government websites, officials' emails, large technology companies, and election infrastructure. The analysis shows that the two media's different ways of framing hackers are underpinned by the **ideologies** behind them and the Chinese and US socio-political landscapes. This study can provide insights into how hacker discourse in media is represented in the 21st century and how national identities are constructed in the media representations of hackers.

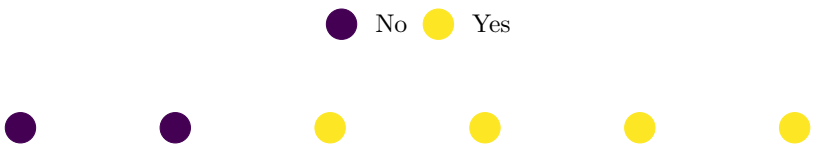


ID: 878

Agreement Score: 0.33

Region-based convolutional neural network using group sparse regularization for image sentiment classification

As an information carrier with rich semantics, images contain more sentiment than texts and audios. So, images are increasingly used by people to express their **opinions** and sentiments in social network. The sentiments of the images are overall and should come from different regions. So, the recognition of the sentiment regions will help to concentrate on important factors the affect the sentiments. Meanwhile, deep learning method for image sentiment classification needs simple and efficient approach for simultaneously carrying out pruning and feature selection whilst optimizing the weights. Motivated by these observations, we design a region-based convolutional neural network using group sparse regularization for image sentiment classification: R-CNNGSR. The method obtains the initial sentiment prediction model through CNN using group sparse regularization to get compact neural network, and then automatically detect the sentiment regions by combining the underlying features and sentimental features. Finally, the whole image and the sentiment region are fused to predict the overall sentiment of the images. Experiment results demonstrate that our proposed R-CNNGSR significantly outperforms the state-of-the-art methods in image sentiment classification.

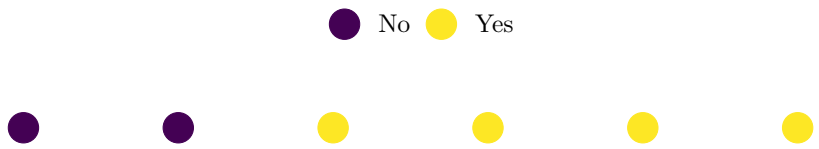


ID: 5922

Agreement Score: 0.33

Classification, detection and sentiment analysis using machine learning over next generation communication platforms

Today, digital news serve a vital function in society by offering a source of news and information to the general public. They serve as a means of communication and can help to educate and inform people about current events and issues. Additionally, newspapers can provide a platform for diverse perspectives and **opinions**, which can help to promote democracy and freedom of expression. Since a large section of the public now turns to internet sources rather than traditional print or broadcast media for news and information, the trend of digital news has gained major support. There are several issues that are commonly associated with digital news. One of the main issues is the proliferation of fake news and misinformation. Additionally, digital news is often subject to algo-rithmically driven personalization, which can lead to a lack of diversity and balance in the information that is presented to people. In this research, we provide an integrated strategy with three key aspects to process and analyse the growing digital news articles. The news data is collected, analysed and classified into different categories using machine learning algorithms. Our news classification results demonstrated that Crime, Cure and Treatment, Economy, Communal, and Entertainment are widely reported news category themes across India. Our Fake News Detection Model achieved 87% accuracy and analysis showed that communal theme is most affected in India. Further sentiment analysis model classified news articles into positive and negative news articles and achieved 89% accuracy.

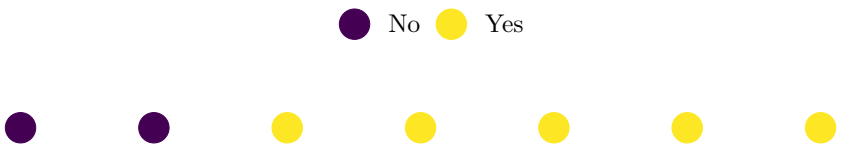


ID: 2858

Agreement Score: 0.33

Extracting and Clustering Main Ideas from Student Feedback Using Language Models

Feedback mechanisms for academic courses have been widely used to measure students **opinions** and satisfaction towards different components of a course; concurrently, open-text detailed impressions enable professors to continually improve their course. However, the process of reading through hundreds of student feedback responses across multiple subjects, followed by the extraction of important ideas is very time consuming. In this work, we propose an automated feedback summarizer to extract the main ideas expressed by all students on various components for each course, based on a pipeline integrating state-of-the-art Natural Language Processing techniques. Our method involves the usage of BERT language models to extract keywords for each course, identify relevant contexts for recurring keywords, and cluster similar contexts. We validate our tool on 8,201 feedback responses for 168 distinct courses from the Computer Science Department of University Politehnica of Bucharest for the 2019-2020 academic year. Our approach achieves a size reduction of 59% on the overall volume of text, while only increasing the mean average error when predicting course ratings from student open-text feedback by an absolute value of 0.06.



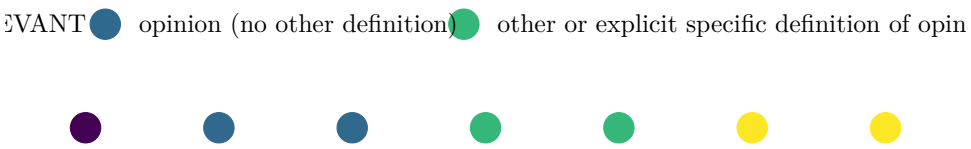
ID: 3913

Agreement Score: 0.43

Trump's Twitter Propaganda During Covid-19

During the early Covid-19 pandemic, led by controversial presidential figure Donald Trump, the US seemed to be overwhelmed by this microbial creature, proving to be one of the countries with the most Covid-19. Besides health impacts, there are many multi-effects afterward, such as economic, social, political, and so on, that must be faced after this Covid-19 outbreak. Moreover, the US will hold a presidential election in November 2020. This challenge makes Trump must focus on how to complete Covid-19 while maintaining electability as President. One of the methods adopted is by forming narration through Twitter. Donald Trump is an active Twitter user who often tweets about his **stance**. Therefore, this study wants to analyze Trump's Twitter tweet content based on propaganda based on six propaganda classifications based on Holly Thayer's Theory. The quantitative content analysis method is the systematic and replicable examination of symbols of communication. The object of research in this article is Twitter's @realDonaldTrump tweet. We analyzed Donald Trump's Twitter content in the period 1 March 2020-27 May 2020 with a systematic random sample method. Our result shows that Trump constructs a message to support his policy and maintain his electability.

2 Concept

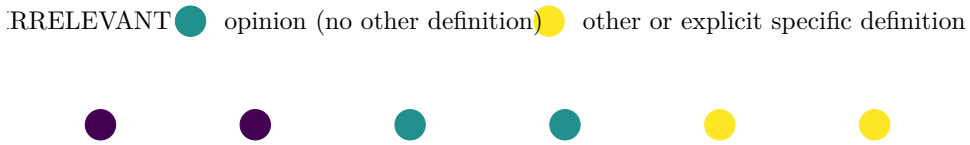


ID: 330

Agreement Score: -0.43

Dynamical mining of ever-changing user requirements: A product design and improvement perspective

Previous studies carried out customer surveys by questionnaires to collect data for analyzing consumer requirements. In recent years, a large and growing body of literature has investigated the extraction of customer requirements and preferences from online reviews. However, since customer requirements change dynamically over time, traditional studies failed to obtain the change data of customer requirements and **opinions** based on sentiments expressed in reviews. In this paper, a new method for dynamically mining user requirements is proposed, which is used to analyze the changing behavior of product attributes and improve product design. Dynamic mining differs from the traditional need acquisition mainly in three aspects: (1) it involves dynamically mining user requirements over time (2) it adds changes in manufacturers' **opinions** to the analysis (3) it allows for product improvement strategies based on the changing behavior of product attributes. First, text mining is adopted to collect customer and manufacturer review data for different time periods and extract product attributes. A Natural Language Processing tool is used to measure the importance weight and sentiment score of product attributes. Second, an approach for dynamically mining user requirements is introduced to classify product attributes and analyze the changes of attribute data in three categories over time. Finally, an improvement strategy for next-generation product design is developed based on the changing behavior of attributes. Moreover, a case study on vehicles based on online reviews was conducted to illustrate the proposed methodology. Our research suggests that the proposed approach can accurately mine customer requirements and lead to successful product improvement strategies for next-generation products.

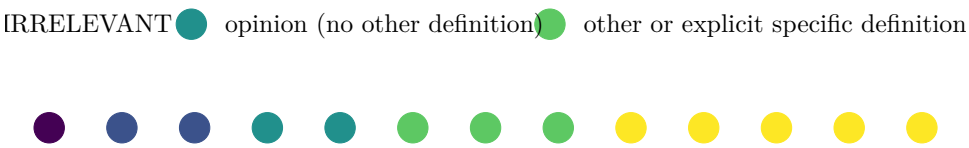


ID: 2858

Agreement Score: -0.33

Extracting and Clustering Main Ideas from Student Feedback Using Language Models

Feedback mechanisms for academic courses have been widely used to measure students **opinions** and satisfaction towards different components of a course; concurrently, open-text detailed impressions enable professors to continually improve their course. However, the process of reading through hundreds of student feedback responses across multiple subjects, followed by the extraction of important ideas is very time consuming. In this work, we propose an automated feedback summarizer to extract the main ideas expressed by all students on various components for each course, based on a pipeline integrating state-of-the-art Natural Language Processing techniques. Our method involves the usage of BERT language models to extract keywords for each course, identify relevant contexts for recurring keywords, and cluster similar contexts. We validate our tool on 8,201 feedback responses for 168 distinct courses from the Computer Science Department of University Politehnica of Bucharest for the 2019-2020 academic year. Our approach achieves a size reduction of 59% on the overall volume of text, while only increasing the mean average error when predicting course ratings from student open-text feedback by an absolute value of 0.06.

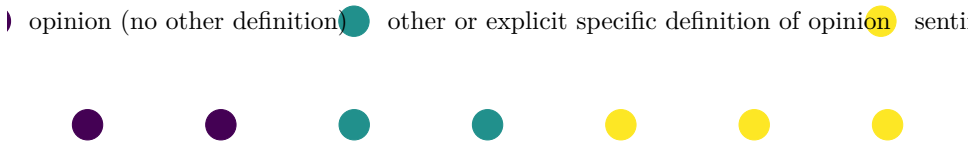


ID: 5026

Agreement Score: -0.23

Public Opinion Mining on Construction Health and Safety: Latent Dirichlet Allocation Approach

The construction industry has been experiencing many occupational accidents as working on construction sites is dangerous. To reduce the likelihood of accidents, construction companies share the latest construction health and safety news and information on social media. While research studies in recent years have explored the perceptions towards these companies' social media pages, there are no big data analytic studies conducted on Instagram about construction health and safety. This study aims to consolidate public perceptions of construction health and safety by analyzing Instagram posts. The study adopted a big data analytics approach involving visual, content, user, and sentiment analyses of Instagram posts (n = 17,835). The study adopted the Latent Dirichlet Allocation, a kind of machine learning approach for generative probabilistic topic extraction, and the five most mentioned topics were: (a) training service, (b) team management, (c) training organization, (d) workers' work and family, and (e) users' action. Besides, the Jaccard coefficient co-occurrence cluster analysis revealed: (a) the most mentioned collocations were 'construction safety week', 'safety first', and 'construction team', (b) the largest clusters were 'safety training', 'occupational health and safety administration', and 'health and safety environment', (c) the most active users were 'Parallel Consultancy Ltd.', 'Pike Consulting Group', and 'Global Training Canada', and (d) positive sentiment accounted for an overwhelming figure of 85%. The findings inform the industry on public perceptions that help create awareness and develop preventative measures for increased health and safety and decreased incidents.

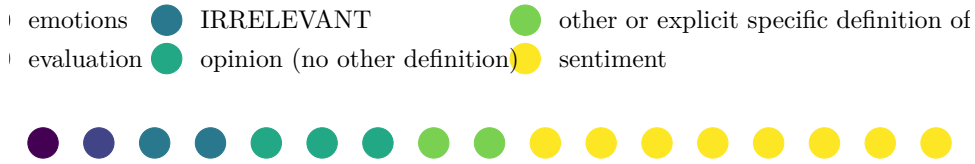


ID: 1749

Agreement Score: -0.14

Analyzing mass media influence using natural language processing and time series analysis

A key question of collective social behavior is related to the influence of mass media on public **opinion**. Different approaches have been developed to address quantitatively this issue, ranging from field experiments to mathematical models. In this work we propose a combination of tools involving natural language processing and time series analysis. We compare selected features of mass media news articles with measurable manifestation of public **opinion**. We apply our analysis to news articles belonging to the 2016 US presidential campaign. We compare variations in polls (as a proxy of public **opinion**) with changes in the connotation of the news (sentiment) or in the agenda (topics) of a selected group of media outlets. Our results suggest that the sentiment content by itself is not enough to understand the differences in polls, but the combination of topics coverage and sentiment content provides an useful insight of the context in which public **opinion** varies. The methodology employed in this work is far general and can be easily extended to other topics of interest.

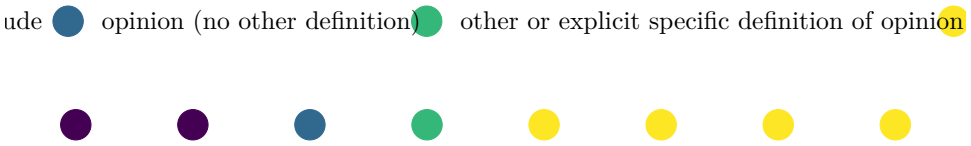


ID: 1017

Agreement Score: -0.06

NLP-Based Customer Loyalty Improvement Recommender System (CLIRS2)

Structured data on customer feedback is becoming more costly and timely to collect and organize. On the other hand, unstructured **opinionated** data, e.g., in the form of free-text comments, is proliferating and available on public websites, such as social media websites, blogs, forums, and websites that provide recommendations. This research proposes a novel method to develop a knowledge-based recommender system from unstructured (text) data. The method is based on applying an **opinion** mining algorithm, extracting aspect-based sentiment score per text item, and transforming text into a structured form. An action rule mining algorithm is applied to the data table constructed from sentiment mining. The proposed application of the method is the problem of improving customer satisfaction ratings. The results obtained from the dataset of customer comments related to the repair services were evaluated with accuracy and coverage. Further, the results were incorporated into the framework of a web-based user-friendly recommender system to advise the business on how to maximally increase their profits by introducing minimal sets of changes in their service. Experiments and evaluation results from comparing the structured data-based version of the system CLIRS (Customer Loyalty Improvement Recommender System) with the unstructured data-based version of the system (CLIRS2) are provided.

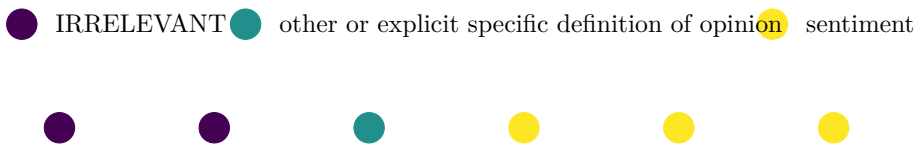


ID: 6070

Agreement Score: 0

Examining the Impact of Discretization Technique on Sentiment Analysis for the Greek Language

Nowadays, information, communication and interaction between people worldwide have been facilitated by the rapid development of technology and they are mainly achieved through the internet. Internet users are now new creators of information data and express their ideas, their **opinions**, their feelings and their **attitudes** about products and services rather than passive information recipients. Given the evolution of modern technological advances, such as the proliferation of mobile devices social networks and services is extending. User-generated content in social media constitutes a very meaningful information source and consists of **opinions** towards various events and services. In this paper, we present a methodology- that aims to analyze Greek text and extract indicative info towards users' **opinions** and **attitudes**. Specifically, we describe a supervised approach adopted that analyzes and classifies comments and reviews into the appropriate polarity category. Discretization techniques are also applied to improve the performance and the accuracy of classification procedures. Finally, we present an experimental evaluation that was designed and conducted and which revealed quite interesting findings.

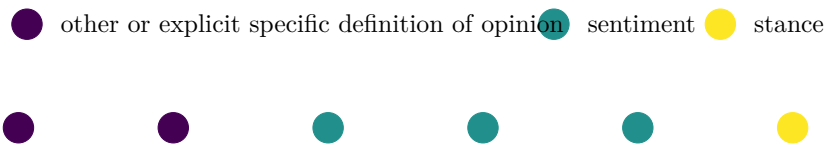


ID: 878

Agreement Score: 0

Region-based convolutional neural network using group sparse regularization for image sentiment classification

As an information carrier with rich semantics, images contain more sentiment than texts and audios. So, images are increasingly used by people to express their **opinions** and sentiments in social network. The sentiments of the images are overall and should come from different regions. So, the recognition of the sentiment regions will help to concentrate on important factors the affect the sentiments. Meanwhile, deep learning method for image sentiment classification needs simple and efficient approach for simultaneously carrying out pruning and feature selection whilst optimizing the weights. Motivated by these observations, we design a region-based convolutional neural network using group sparse regularization for image sentiment classification: R-CNNGSR. The method obtains the initial sentiment prediction model through CNN using group sparse regularization to get compact neural network, and then automatically detect the sentiment regions by combining the underlying features and sentimental features. Finally, the whole image and the sentiment region are fused to predict the overall sentiment of the images. Experiment results demonstrate that our proposed R-CNNGSR significantly outperforms the state-of-the-art methods in image sentiment classification.

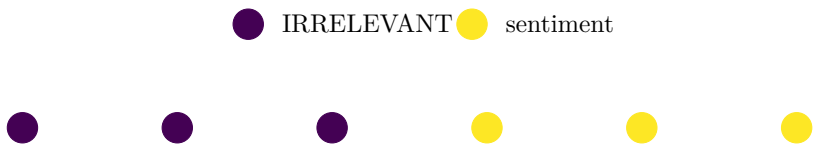


ID: 2849

Agreement Score: 0

A new dataset of Dutch and Danish party congress speeches

We present a new dataset of speeches given by Danish and Dutch politicians at party congresses between 1946 and 2017. The dataset is a unique collection of materials from different party archives and digital repositories. It offers a unique opportunity to analyse the issues discussed in these speeches, the **positions** taken and the rhetoric used by party elites over time and between countries. We describe the data and illustrate them with one application: a sentiment analysis that describes differences between parties and over time.

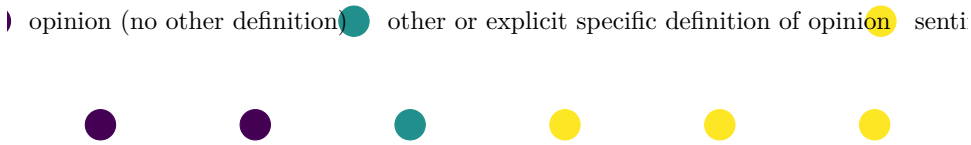


ID: 2378

Agreement Score: 0

Towards a Service-Oriented Architecture for Pre-processing Crowd-Sourced Sentiment from Twitter

Online social media platforms like Twitter, provide **opinion** rich repositories for conducting sentiment analysis. Users engage in open discussions on a variety of topics across a wide cross-section of problem domains. Commercial, government, educational, non-profit and other types of agencies are increasingly relying on extracting conversations on Twitter to determine the general sentiment of the public on particular topics, products, services and issues. Despite being readily available and in abundance, it is also laced with nuances which can disrupt, skew and potentially lead to inaccurate analysis if not handled properly. In this paper, we propose an SOA framework to enable the pre-processing of data origination on Twitter, and configurable components that allow data consumers to filter the data using useful social media signals.

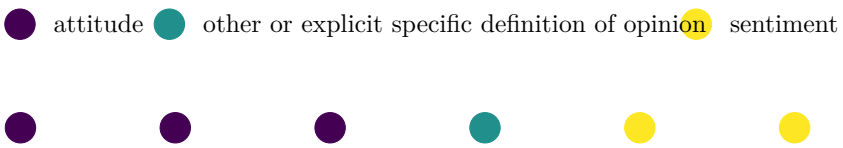


ID: 4633

Agreement Score: 0

Opinion Mining of Readers' Responses to Literary Prize Nominees on Twitter: A Case Study of Public Reaction to the Booker Prize (2018-2020)

The award of literary prizes such as the Booker Prize has a great impact on production and reception of literary works. In recent years, the Booker Prize awarding committee has faced challenges in selecting books that are more readable and popular. This study suggests that one feasible way to address this issue is to analyze readers' response and public reaction to literary works in social media. In this regard, the present study aims to develop a data analytics technique that can analyze literary readers' responses in Twitter and predict a potential prize winner. To achieve this focal goal, a Sentiment Analysis and Topic Modelling approach is designed to classify the public reactions to The Booker Prize nominees in Twitter. The data is extracted for three consecutive years of 2018, 2019 and 2020. In addition, the study utilizes Machine Learning to propose a prediction technique in selecting the best possible Booker Prize winner based on public **opinion**. The results reveal the main topics frequently appearing in the data as well as the positive and negative sentiments attached to them.

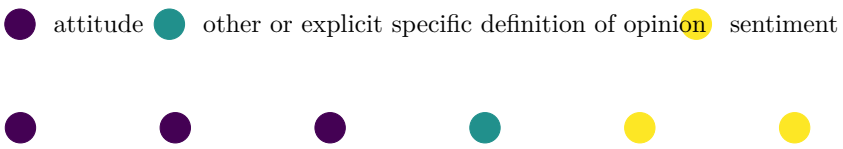


ID: 4344

Agreement Score: 0

Sentiment Analysis on Multimodal Transportation during the COVID-19 Using Social Media Data

This paper aims to leverage Twitter data to understand travel mode choices during the pandemic. Tweets related to different travel modes in New York City (NYC) are fetched from Twitter in the two most recent years (January 2020-January 2022). Building on these data, we develop travel mode classifiers, adapted from natural language processing (NLP) models, to determine whether individual tweets are related to some travel mode (subway, bus, bike, taxi/Uber, and private vehicle). Sentiment analysis is performed to understand people's **attitudinal** changes about mode choices during the pandemic. Results show that a majority of people had a positive **attitude** toward buses, bikes, and private vehicles, which is consistent with the phenomenon of many commuters shifting away from subways to buses, bikes and private vehicles during the pandemic. We analyze negative tweets related to travel modes and find that people were worried about those who did not wear masks on subways and buses. Based on users' demographic information, we conduct regression analysis to analyze what factors affected people's **attitude** toward public transit. We find that the **attitude** of users in the service industry was more easily affected by MTA subway service during the pandemic.

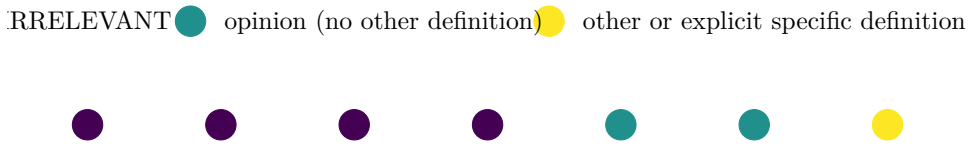


ID: 2426

Agreement Score: 0

Deep Learning for Information Triage on Twitter

In this paper, we present a Deep Learning-based system for the support of information triaging on Twitter during emergency situations, such as disasters, or other influential events, such as political elections. The system is based on the assumption that a different type of information is required right after the event and some time after the event occurs. In a preliminary study, we analyze the language behavior of Twitter users during two kinds of influential events, namely, natural disasters and political elections. In the study, we analyze the credibility of information included by users in tweets in the above-mentioned situations, by classifying the information into two kinds: Primary Information (first-hand reports) and Secondary Information (second-hand reports, retweets, etc.). We also perform sentiment analysis of the data to check user **attitudes** toward the occurring events. Next, we present the structure of the system and compare a number of classifiers, including the proposed one based on Convolutional Neural Networks. Finally, we validate the system by performing an in-depth analysis of information obtained after a number of additional events, including an eruption of a Japanese volcano Ontake on 27 September 2014, as well as heavy rains and typhoons that occurred in 2020. We confirm that the methods works sufficiently well even when trained on data from nearly 10 years ago, which strongly suggests that the model is well-generalized and sufficiently grasps important aspects of each type of classified information.

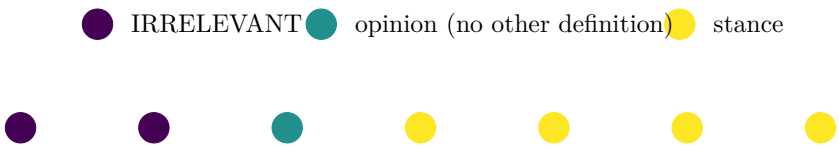


ID: 1799

Agreement Score: 0.14

Review article: Detection of actionable tweets in crisis events

Messages on social media can be an important source of information during crisis situations. They can frequently provide details about developments much faster than traditional sources (e.g., official news) and can offer personal perspectives on events, such as **opinions** or specific needs. In the future, these messages can also serve to assess disaster risks. One challenge for utilizing social media in crisis situations is the reliable detection of relevant messages in a flood of data. Researchers have started to look into this problem in recent years, beginning with crowdsourced methods. Lately, approaches have shifted towards an automatic analysis of messages. A major stumbling block here is the question of exactly what messages are considered relevant or informative, as this is dependent on the specific usage scenario and the role of the user in this scenario. In this review article, we present methods for the automatic detection of crisis-related messages (tweets) on Twitter. We start by showing the varying definitions of importance and relevance relating to disasters, leading into the concept of use case-dependent actionability that has recently become more popular and is the focal point of the review paper. This is followed by an overview of existing crisis-related social media data sets for evaluation and training purposes. We then compare approaches for solving the detection problem based (1) on filtering by characteristics like keywords and location, (2) on crowdsourcing, and (3) on machine learning technique. We analyze their suitability and limitations of the approaches with regards to actionability. We then point out particular challenges, such as the linguistic issues concerning social media data. Finally, we suggest future avenues of research and show connections to related tasks, such as the subsequent semantic classification of tweets.

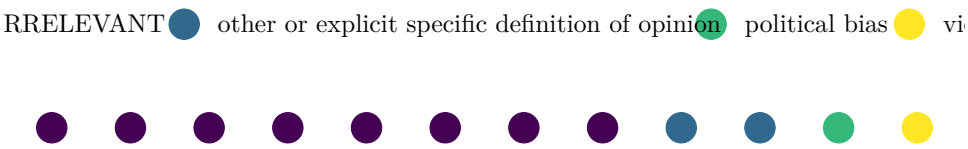


ID: 3913

Agreement Score: 0.14

Trump's Twitter Propaganda During Covid-19

During the early Covid-19 pandemic, led by controversial presidential figure Donald Trump, the US seemed to be overwhelmed by this microbial creature, proving to be one of the countries with the most Covid-19. Besides health impacts, there are many multi-effects afterward, such as economic, social, political, and so on, that must be faced after this Covid-19 outbreak. Moreover, the US will hold a presidential election in November 2020. This challenge makes Trump must focus on how to complete Covid-19 while maintaining electability as President. One of the methods adopted is by forming narration through Twitter. Donald Trump is an active Twitter user who often tweets about his **stance**. Therefore, this study wants to analyze Trump's Twitter tweet content based on propaganda based on six propaganda classifications based on Holly Thayer's Theory. The quantitative content analysis method is the systematic and replicable examination of symbols of communication. The object of research in this article is Twitter's @realDonaldTrump tweet. We analyzed Donald Trump's Twitter content in the period 1 March 2020-27 May 2020 with a systematic random sample method. Our result shows that Trump constructs a message to support his policy and maintain his electability.

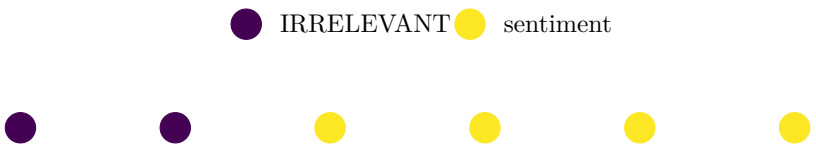


ID: 1533

Agreement Score: 0.33

Media portrayal of hackers in China Daily and The New York Times: A corpus-based critical discourse analysis

This study draws on a synergy of Corpus Linguistics and Critical Discourse Studies to scrutinize the portrayal of hackers in China Daily and The New York Times in the 21st century (2001-2020), primarily revolving around the main social actors and targets in hacking. This study demonstrates that both media share a positive transformation of the image-building of hackers in the 21st century. Besides, countries are salient social actors in hacker media discourse and the two media differ in their ways of constructing them. The New York Times tends to have a negative other-representation and categorical otherness of specific countries through such discursive strategies as negative other-representation and group categorization, whereas China Daily is prone to insist on opposing the US hacking allegations in a defensive manner. Regarding major targets, China Daily highlights government websites whereas The New York Times emphasizes government websites, officials' emails, large technology companies, and election infrastructure. The analysis shows that the two media's different ways of framing hackers are underpinned by the **ideologies** behind them and the Chinese and US socio-political landscapes. This study can provide insights into how hacker discourse in media is represented in the 21st century and how national identities are constructed in the media representations of hackers.

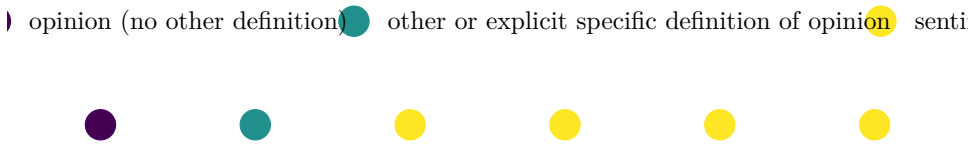


ID: 5922

Agreement Score: 0.33

Classification, detection and sentiment analysis using machine learning over next generation communication platforms

Today, digital news serve a vital function in society by offering a source of news and information to the general public. They serve as a means of communication and can help to educate and inform people about current events and issues. Additionally, newspapers can provide a platform for diverse perspectives and **opinions**, which can help to promote democracy and freedom of expression. Since a large section of the public now turns to internet sources rather than traditional print or broadcast media for news and information, the trend of digital news has gained major support. There are several issues that are commonly associated with digital news. One of the main issues is the proliferation of fake news and misinformation. Additionally, digital news is often subject to algo-rithmically driven personalization, which can lead to a lack of diversity and balance in the information that is presented to people. In this research, we provide an integrated strategy with three key aspects to process and analyse the growing digital news articles. The news data is collected, analysed and classified into different categories using machine learning algorithms. Our news classification results demonstrated that Crime, Cure and Treatment, Economy, Communal, and Entertainment are widely reported news category themes across India. Our Fake News Detection Model achieved 87% accuracy and analysis showed that communal theme is most affected in India. Further sentiment analysis model classified news articles into positive and negative news articles and achieved 89% accuracy.

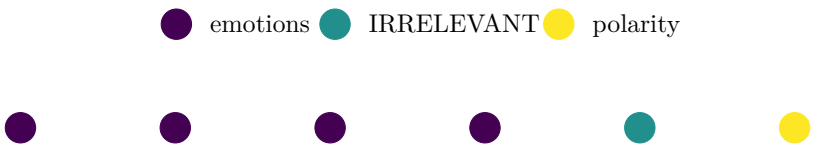


ID: 5071

Agreement Score: 0.33

Comparison Performance of Long Short-Term Memory and Convolution Neural Network Variants on Online Learning Tweet Sentiment Analysis

Sentiment analysis can be act as an assisted tool in improving the quality of online teaching and learning between teachers and students. Twitter social media platform currently more than 500 million tweets sent each day which is equal to 5787 tweets per second. Therefore, it is hard to track users' overall **opinions** on the topics contained in social media. To catch up with the feedback on online learning, it is crucial to detect the topic being discussed and classify users' sentiments towards those topics. Even though there are many approaches in developing sentiment analysis models, DL models prove to provide the best performance in the sentiment analysis field. Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) are two mainstream models in DL used for sentiment analysis classification. Therefore, we evaluate CNN, LSTM, and its hybrids to classify sentiment or an online learning tweet from 2020 until 2021 of 23168 tweets. CNN-LSTM, LSTM-CNN, Bidirectional LSTM, CNN-Bidirectional LSTM models were designed and evaluated based on random hyper-parameter tuning. We explain the proposed methodology and model design illustration. The outcome assesses the superiority of all models with a remarkable improvement of accuracy and a reduction loss when applying the random oversampling technique. Specifically, the LSTM-CNN model with random oversampling technique outperformed the other six models with an accuracy of 87.40% and loss value of 0.3432. However, the computational time has resulted increased when with random oversampling technique. Thus, in the future, the performance can be improved on computational time and hyperparameter selection with the employment of nature-inspired computing for fast and optimal results.

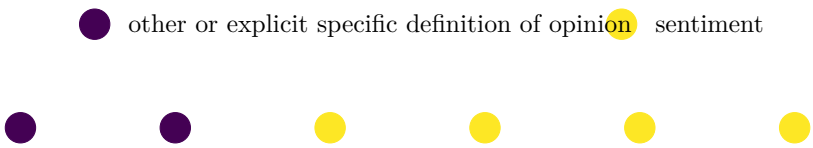


ID: 2087

Agreement Score: 0.33

Adaptive Learning Emotion Identification Method of Short Texts for Online Medical Knowledge Sharing Community

The medical knowledge sharing community provides users with an open platform for accessing medical resources and sharing medical knowledge, treatment experience, and emotions. Compared with the recipients of general commodities, the recipients in the medical knowledge sharing community pay more attention to the intensity or overall evaluation of emotional vocabularies in the comments, such as treatment effects, prices, service **attitudes**, and other aspects. Therefore, the overall evaluation is not a key factor in medical service comments, but the semantics of the emotional polarity is the key to affect recipients of the medical information. In this paper, we propose an adaptive learning emotion identification method (ALEIM) based on mutual information feature weight, which captures the correlation and redundancy of features. In order to evaluate the proposed method's effectiveness, we use four basic corpus libraries crawled from the Haodf's online platform and employ Taiwan University NTUSD Simplified Chinese Emotion Dictionary for emotion classification. The experimental results show that our proposed ALEIM method has a better performance for the identification of the low-frequency words' redundant features in comments of the online medical knowledge sharing community.

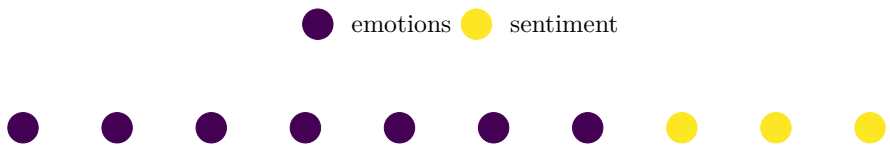


ID: 1696

Agreement Score: 0.33

Knowledge-Guided Sentiment Analysis Via Learning From Natural Language Explanations

Sentiment analysis is crucial for studying public **opinion** since it can provide us with valuable information. Existing sentiment analysis methods rely on finding the sentiment element from the content of user-generated. However, the question of why a message produces certain emotions has not been well explored or utilized in previous works. To address this challenge, we propose a natural language explanation framework for sentiment analysis that provides sufficient domain knowledge for generating additional labelled data for each new labelling decision. A rule-based semantic parser transforms these explanations into programmatic labelling functions that generate noisy labels for an arbitrary amount of unlabelled sentiment information to train a sentiment analysis classifier. Experiments on two sentiment analysis datasets demonstrate the superiority it achieves over baseline methods by leveraging explanations as external knowledge to joint training a sentiment analysis model rather than only labels. An ablation study is conducted to clarify the relative contribution of natural language explanations.



ID: 4065

Agreement Score: 0.4

Enhancing Digital Well-being using Opinion Mining and Sentiment Classifiers

Opinion mining on various issues is a very popular trend in the micro-blogging research community. Advanced data mining techniques using sentiment analysis and machine learning algorithms on large datasets like microblogging websites are popular and trending in the data science community. But such analytics are limited to only certain aspects of interest. In this paper, we present a search engine that presents a user's emotions over a timeline. This would serve as a novel approach to adding search capability over **opinion** mining procedures and exploring potential optimizations for a wide range of features and methods for training sentiment classifiers. Users should be able to chart down their emotional quotient using the proposed application. This would aid in promoting their digital well-being.

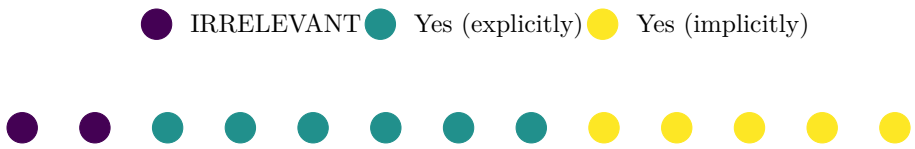
3 relevant-2

ID: 1017

Agreement Score: -0.18

NLP-Based Customer Loyalty Improvement Recommender System (CLIRS2)

Structured data on customer feedback is becoming more costly and timely to collect and organize. On the other hand, unstructured **opinionated** data, e.g., in the form of free-text comments, is proliferating and available on public websites, such as social media websites, blogs, forums, and websites that provide recommendations. This research proposes a novel method to develop a knowledge-based recommender system from unstructured (text) data. The method is based on applying an **opinion** mining algorithm, extracting aspect-based sentiment score per text item, and transforming text into a structured form. An action rule mining algorithm is applied to the data table constructed from sentiment mining. The proposed application of the method is the problem of improving customer satisfaction ratings. The results obtained from the dataset of customer comments related to the repair services were evaluated with accuracy and coverage. Further, the results were incorporated into the framework of a web-based user-friendly recommender system to advise the business on how to maximally increase their profits by introducing minimal sets of changes in their service. Experiments and evaluation results from comparing the structured data-based version of the system CLIRS (Customer Loyalty Improvement Recommender System) with the unstructured data-based version of the system (CLIRS2) are provided.

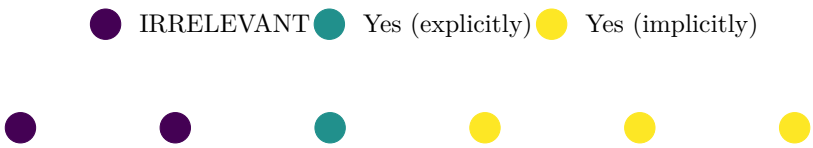


ID: 5026

Agreement Score: -0.08

Public Opinion Mining on Construction Health and Safety: Latent Dirichlet Allocation Approach

The construction industry has been experiencing many occupational accidents as working on construction sites is dangerous. To reduce the likelihood of accidents, construction companies share the latest construction health and safety news and information on social media. While research studies in recent years have explored the perceptions towards these companies' social media pages, there are no big data analytic studies conducted on Instagram about construction health and safety. This study aims to consolidate public perceptions of construction health and safety by analyzing Instagram posts. The study adopted a big data analytics approach involving visual, content, user, and sentiment analyses of Instagram posts (n = 17,835). The study adopted the Latent Dirichlet Allocation, a kind of machine learning approach for generative probabilistic topic extraction, and the five most mentioned topics were: (a) training service, (b) team management, (c) training organization, (d) workers' work and family, and (e) users' action. Besides, the Jaccard coefficient co-occurrence cluster analysis revealed: (a) the most mentioned collocations were 'construction safety week', 'safety first', and 'construction team', (b) the largest clusters were 'safety training', 'occupational health and safety administration', and 'health and safety environment', (c) the most active users were 'Parallel Consultancy Ltd.', 'Pike Consulting Group', and 'Global Training Canada', and (d) positive sentiment accounted for an overwhelming figure of 85%. The findings inform the industry on public perceptions that help create awareness and develop preventative measures for increased health and safety and decreased incidents.

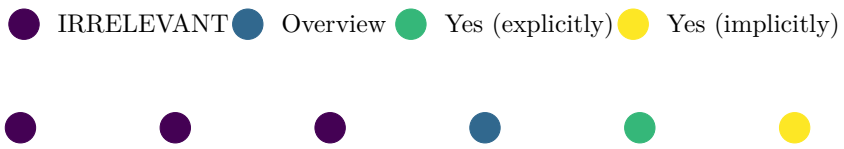


ID: 5922

Agreement Score: 0

Classification, detection and sentiment analysis using machine learning over next generation communication platforms

Today, digital news serve a vital function in society by offering a source of news and information to the general public. They serve as a means of communication and can help to educate and inform people about current events and issues. Additionally, newspapers can provide a platform for diverse perspectives and **opinions**, which can help to promote democracy and freedom of expression. Since a large section of the public now turns to internet sources rather than traditional print or broadcast media for news and information, the trend of digital news has gained major support. There are several issues that are commonly associated with digital news. One of the main issues is the proliferation of fake news and misinformation. Additionally, digital news is often subject to algo-rithmically driven personalization, which can lead to a lack of diversity and balance in the information that is presented to people. In this research, we provide an integrated strategy with three key aspects to process and analyse the growing digital news articles. The news data is collected, analysed and classified into different categories using machine learning algorithms. Our news classification results demonstrated that Crime, Cure and Treatment, Economy, Communal, and Entertainment are widely reported news category themes across India. Our Fake News Detection Model achieved 87% accuracy and analysis showed that communal theme is most affected in India. Further sentiment analysis model classified news articles into positive and negative news articles and achieved 89% accuracy.

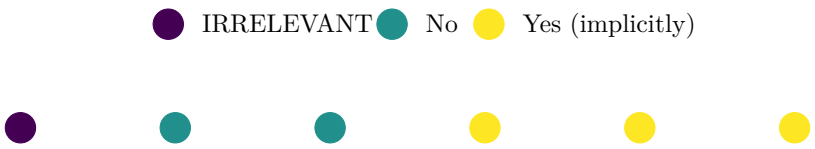


ID: 2378

Agreement Score: 0

Towards a Service-Oriented Architecture for Pre-processing Crowd-Sourced Sentiment from Twitter

Online social media platforms like Twitter, provide **opinion** rich repositories for conducting sentiment analysis. Users engage in open discussions on a variety of topics across a wide cross-section of problem domains. Commercial, government, educational, non-profit and other types of agencies are increasingly relying on extracting conversations on Twitter to determine the general sentiment of the public on particular topics, products, services and issues. Despite being readily available and in abundance, it is also laced with nuances which can disrupt, skew and potentially lead to inaccurate analysis if not handled properly. In this paper, we propose an SOA framework to enable the pre-processing of data origination on Twitter, and configurable components that allow data consumers to filter the data using useful social media signals.

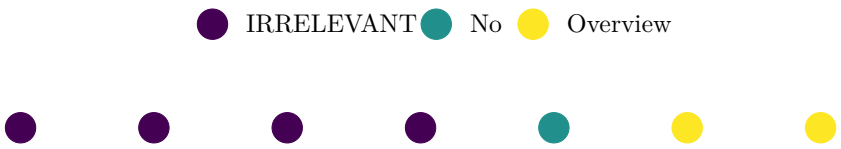


ID: 1530

Agreement Score: 0

Opinion Prediction of Hungarian Students for Real-Time E-Learning Systems: A Futuristic Sustainable Technology-Based Solution

This work is a new step towards the understanding of students' **opinions** about the use of technology in learning and improvements to provide sustainable E-learning solutions. Every higher educational university tries to provide well-suited, updated, and trending technology-based education facilities to its students. The task of analyzing the student's sentiment about technology delivers benefits not only to ICT administrators, but also to management to become aware of the technological concerns. The **opinions** of Hungarian university students were analyzed using the regression method. We investigated 165 primary samples supported by the four hypotheses. The reliability of the data sample was calculated as 0.91 with Cronbach alpha testing. The Pearson Momentum Correlation (PMC) proved that the suggested technology benefits had a linear positive association with the student's **opinion**. Furthermore, technology usability was positively correlated with the benefits. The supporting results of the regression model evidenced the significant impact of technology usability and benefits on the **opinions**. Using Exploratory Factor Analysis (EFA), we proposed significant features for the model that predicted students' **opinions** using the educational benefit and usability parameters. These parameters statistically significantly predicted student's **opinions**: $F\ (2,\ 162) = 104.9, p < 0.05, R2 = 0.559$. This study may be supportive of implementing the **opinion** mining model online and useful to university authorities to understand better the students' sentiments about the current technological facilities provided. The authors proposed an **opinion** mining model to deploy on the university's real-time E-lection sustainable technology.



ID: 1799

Agreement Score: 0.14

Review article: Detection of actionable tweets in crisis events

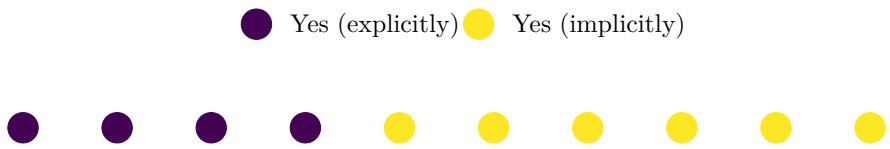
Messages on social media can be an important source of information during crisis situations. They can frequently provide details about developments much faster than traditional sources (e.g., official news) and can offer personal perspectives on events, such as **opinions** or specific needs. In the future, these messages can also serve to assess disaster risks. One challenge for utilizing social media in crisis situations is the reliable detection of relevant messages in a flood of data. Researchers have started to look into this problem in recent years, beginning with crowdsourced methods. Lately, approaches have shifted towards an automatic analysis of messages. A major stumbling block here is the question of exactly what messages are considered relevant or informative, as this is dependent on the specific usage scenario and the role of the user in this scenario. In this review article, we present methods for the automatic detection of crisis-related messages (tweets) on Twitter. We start by showing the varying definitions of importance and relevance relating to disasters, leading into the concept of use case-dependent actionability that has recently become more popular and is the focal point of the review paper. This is followed by an overview of existing crisis-related social media data sets for evaluation and training purposes. We then compare approaches for solving the detection problem based (1) on filtering by characteristics like keywords and location, (2) on crowdsourcing, and (3) on machine learning technique. We analyze their suitability and limitations of the approaches with regards to actionability. We then point out particular challenges, such as the linguistic issues concerning social media data. Finally, we suggest future avenues of research and show connections to related tasks, such as the subsequent semantic classification of tweets.

ID: 3913

Agreement Score: 0.14

Trump's Twitter Propaganda During Covid-19

During the early Covid-19 pandemic, led by controversial presidential figure Donald Trump, the US seemed to be overwhelmed by this microbial creature, proving to be one of the countries with the most Covid-19. Besides health impacts, there are many multi-effects afterward, such as economic, social, political, and so on, that must be faced after this Covid-19 outbreak. Moreover, the US will hold a presidential election in November 2020. This challenge makes Trump must focus on how to complete Covid-19 while maintaining electability as President. One of the methods adopted is by forming narration through Twitter. Donald Trump is an active Twitter user who often tweets about his **stance**. Therefore, this study wants to analyze Trump's Twitter tweet content based on propaganda based on six propaganda classifications based on Holly Thayer's Theory. The quantitative content analysis method is the systematic and replicable examination of symbols of communication. The object of research in this article is Twitter's @realDonaldTrump tweet. We analyzed Donald Trump's Twitter content in the period 1 March 2020-27 May 2020 with a systematic random sample method. Our result shows that Trump constructs a message to support his policy and maintain his electability.

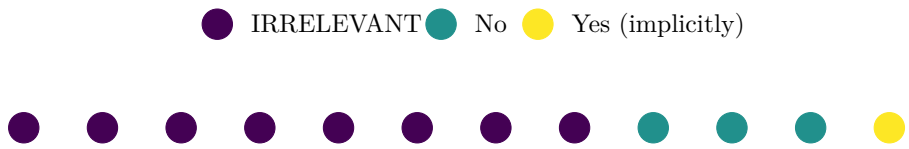


ID: 597

Agreement Score: 0.2

From Stances' Imbalance to Their Hierarchical Representation and Detection

Stance detection has gained increasing interest from the research community due to its importance for fake news detection. The goal of **stance** detection is to categorize an overall **position** of a subject towards an object into one of the four classes: agree, disagree, discuss, and unrelated. One of the major problems faced by current machine learning models used for **stance** detection is caused by a severe class imbalance among these classes. Hence, most models fail to correctly classify instances that fall into minority classes. In this paper, we address this problem by proposing a hierarchical representation of these classes, which combines the agree, disagree, and discuss classes under a new related class. Further, we propose a two-layer neural network that learns from this hierarchical representation and controls the error propagation between the two layers using the Maximum Mean Discrepancy regularizer. Compared with conventional four-way classifiers, this model has two advantages: (1) the hierarchical architecture mitigates the class imbalance problem; (2) the regularization makes the model to better discern between the related and unrelated **stances**. An extensive experimentation demonstrates state-of-the-art accuracy performance of the proposed model for **stance** detection.

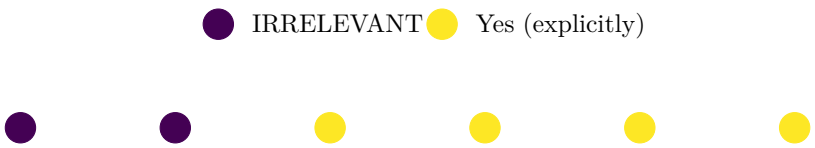


ID: 1533

Agreement Score: 0.33

Media portrayal of hackers in China Daily and The New York Times: A corpus-based critical discourse analysis

This study draws on a synergy of Corpus Linguistics and Critical Discourse Studies to scrutinize the portrayal of hackers in China Daily and The New York Times in the 21st century (2001-2020), primarily revolving around the main social actors and targets in hacking. This study demonstrates that both media share a positive transformation of the image-building of hackers in the 21st century. Besides, countries are salient social actors in hacker media discourse and the two media differ in their ways of constructing them. The New York Times tends to have a negative other-representation and categorical otherness of specific countries through such discursive strategies as negative other-representation and group categorization, whereas China Daily is prone to insist on opposing the US hacking allegations in a defensive manner. Regarding major targets, China Daily highlights government websites whereas The New York Times emphasizes government websites, officials' emails, large technology companies, and election infrastructure. The analysis shows that the two media's different ways of framing hackers are underpinned by the **ideologies** behind them and the Chinese and US socio-political landscapes. This study can provide insights into how hacker discourse in media is represented in the 21st century and how national identities are constructed in the media representations of hackers.

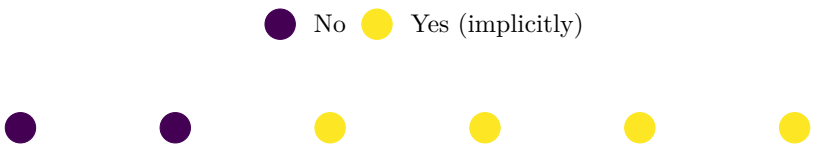


ID: 878

Agreement Score: 0.33

Region-based convolutional neural network using group sparse regularization for image sentiment classification

As an information carrier with rich semantics, images contain more sentiment than texts and audios. So, images are increasingly used by people to express their **opinions** and sentiments in social network. The sentiments of the images are overall and should come from different regions. So, the recognition of the sentiment regions will help to concentrate on important factors the affect the sentiments. Meanwhile, deep learning method for image sentiment classification needs simple and efficient approach for simultaneously carrying out pruning and feature selection whilst optimizing the weights. Motivated by these observations, we design a region-based convolutional neural network using group sparse regularization for image sentiment classification: R-CNNGSR. The method obtains the initial sentiment prediction model through CNN using group sparse regularization to get compact neural network, and then automatically detect the sentiment regions by combining the underlying features and sentimental features. Finally, the whole image and the sentiment region are fused to predict the overall sentiment of the images. Experiment results demonstrate that our proposed R-CNNGSR significantly outperforms the state-of-the-art methods in image sentiment classification.

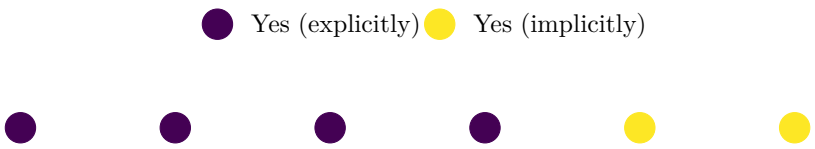


ID: 2849

Agreement Score: 0.33

A new dataset of Dutch and Danish party congress speeches

We present a new dataset of speeches given by Danish and Dutch politicians at party congresses between 1946 and 2017. The dataset is a unique collection of materials from different party archives and digital repositories. It offers a unique opportunity to analyse the issues discussed in these speeches, the **positions** taken and the rhetoric used by party elites over time and between countries. We describe the data and illustrate them with one application: a sentiment analysis that describes differences between parties and over time.

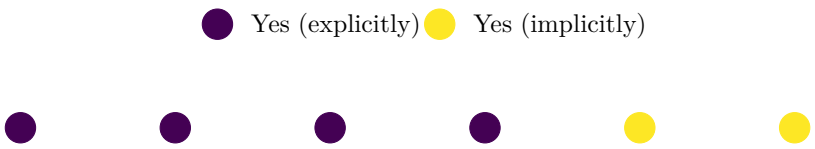


ID: 1446

Agreement Score: 0.33

AB-LaBSE: Uyghur Sentiment Analysis via the Pre-Training Model with BiLSTM

In recent years, more and more attention has been paid to text sentiment analysis, which has gradually become a research hotspot in information extraction, data mining, Natural Language Processing (NLP), and other fields. With the gradual popularization of the Internet, sentiment analysis of Uyghur texts has great research and application value in online public **opinion**. For low-resource languages, most state-of-the-art systems require tens of thousands of annotated sentences to get high performance. However, there is minimal annotated data available about Uyghur sentiment analysis tasks. There are also specificities in each task-differences in words and word order across languages make it a challenging problem. In this paper, we present an effective solution to providing a meaningful and easy-to-use feature extractor for sentiment analysis tasks: using the pre-trained language model with BiLSTM layer. Firstly, data augmentation is carried out by AEDA (An Easier Data Augmentation), and the augmented dataset is constructed to improve the performance of text classification tasks. Then, a pretraining model LaBSE is used to encode the input data. Then, BiLSTM is used to learn more context information. Finally, the validity of the model is verified via two categories datasets for sentiment analysis and five categories datasets for emotion analysis. We evaluated our approach on two datasets, which showed wonderful performance compared to some strong baselines. We close with an overview of the resources for sentiment analysis tasks and some of the open research questions. Therefore, we propose a combined deep learning and cross-language pretraining model for two low resource expectations.

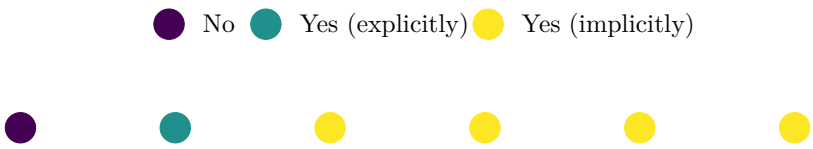


ID: 2159

Agreement Score: 0.33

An Enhanced Sentiment Analysis Framework Based on Pre-Trained Word Embedding

Sentiment analysis (SA) is a technique that lets people in different fields such as business, economy, research, government, and politics to know about people's **opinions**, which greatly affects the process of decision-making. SA techniques are classified into: lexicon-based techniques, machine learning techniques, and a hybrid between both approaches. Each approach has its limitations and drawbacks, the machine learning approach depends on manual feature extraction, lexicon-based approach relies on sentiment lexicons that are usually unscalable, unreliable, and manually annotated by human experts. Nowadays, word-embedding techniques have been commonly used in SA classification. Currently, Word2Vec and GloVe are some of the most accurate and usable word embedding techniques, which can transform words into meaningful semantic vectors. However, these techniques ignore sentiment information of texts and require a huge corpus of texts for training and generating accurate vectors, which are used as inputs of deep learning models. In this paper, we propose an enhanced ensemble classifier framework. Our framework is based on our previously published lexicon-based method, bag-of-words, and pre-trained word embedding, first the sentence is preprocessed by removing stop-words, POS tagging, stemming and lemmatization, shortening exaggerated word. Second, the processed sentence is passed to three modules, our previous lexicon-based method (Sum Votes), bag-of-words module and semantic module (Word2Vec and Glove) and produced feature vectors. Finally, the previous features vectors are fed into 11 different classifiers. The proposed framework is tested and evaluated over four datasets with five different lexicons, the experiment results show that our proposed model outperforms the previous lexicon based and the machine learning methods individually.

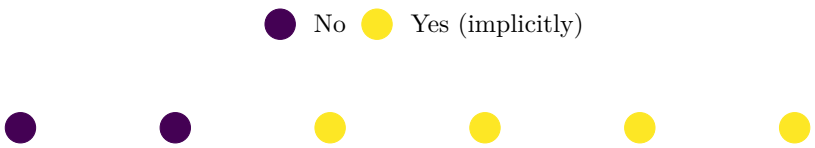


ID: 4136

Agreement Score: 0.33

Variational Weakly Supervised Sentiment Analysis with Posterior Regularization

Sentiment analysis is an important task in natural language processing (NLP). Most of existing state-of-the-art methods are under the supervised learning paradigm. However, human annotations can be scarce. Thus, we should leverage more weak supervision for sentiment analysis. In this paper, we propose a posterior regularization framework for the variational approach to the weakly supervised sentiment analysis to better control the posterior distribution of the label assignment. The intuition behind the posterior regularization is that if extracted **opinion** words from two documents are semantically similar, the posterior distributions of two documents should be similar. Our experimental results show that the posterior regularization can improve the original variational approach to the weakly supervised sentiment analysis and the performance is more stable with smaller prediction variance.

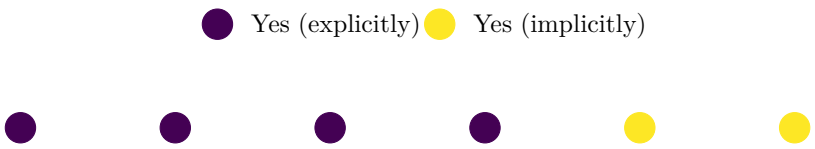


ID: 4633

Agreement Score: 0.33

Opinion Mining of Readers' Responses to Literary Prize Nominees on Twitter: A Case Study of Public Reaction to the Booker Prize (2018-2020)

The award of literary prizes such as the Booker Prize has a great impact on production and reception of literary works. In recent years, the Booker Prize awarding committee has faced challenges in selecting books that are more readable and popular. This study suggests that one feasible way to address this issue is to analyze readers' response and public reaction to literary works in social media. In this regard, the present study aims to develop a data analytics technique that can analyze literary readers' responses in Twitter and predict a potential prize winner. To achieve this focal goal, a Sentiment Analysis and Topic Modelling approach is designed to classify the public reactions to The Booker Prize nominees in Twitter. The data is extracted for three consecutive years of 2018, 2019 and 2020. In addition, the study utilizes Machine Learning to propose a prediction technique in selecting the best possible Booker Prize winner based on public **opinion**. The results reveal the main topics frequently appearing in the data as well as the positive and negative sentiments attached to them.

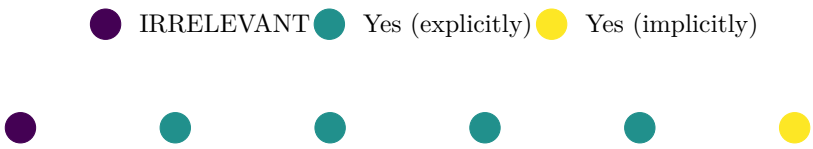


ID: 2426

Agreement Score: 0.33

Deep Learning for Information Triage on Twitter

In this paper, we present a Deep Learning-based system for the support of information triaging on Twitter during emergency situations, such as disasters, or other influential events, such as political elections. The system is based on the assumption that a different type of information is required right after the event and some time after the event occurs. In a preliminary study, we analyze the language behavior of Twitter users during two kinds of influential events, namely, natural disasters and political elections. In the study, we analyze the credibility of information included by users in tweets in the above-mentioned situations, by classifying the information into two kinds: Primary Information (first-hand reports) and Secondary Information (second-hand reports, retweets, etc.). We also perform sentiment analysis of the data to check user **attitudes** toward the occurring events. Next, we present the structure of the system and compare a number of classifiers, including the proposed one based on Convolutional Neural Networks. Finally, we validate the system by performing an in-depth analysis of information obtained after a number of additional events, including an eruption of a Japanese volcano Ontake on 27 September 2014, as well as heavy rains and typhoons that occurred in 2020. We confirm that the methods works sufficiently well even when trained on data from nearly 10 years ago, which strongly suggests that the model is well-generalized and sufficiently grasps important aspects of each type of classified information.

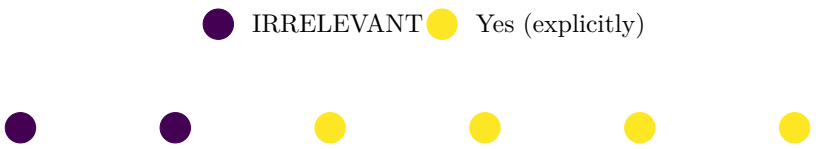


ID: 2087

Agreement Score: 0.33

Adaptive Learning Emotion Identification Method of Short Texts for Online Medical Knowledge Sharing Community

The medical knowledge sharing community provides users with an open platform for accessing medical resources and sharing medical knowledge, treatment experience, and emotions. Compared with the recipients of general commodities, the recipients in the medical knowledge sharing community pay more attention to the intensity or overall evaluation of emotional vocabularies in the comments, such as treatment effects, prices, service **attitudes**, and other aspects. Therefore, the overall evaluation is not a key factor in medical service comments, but the semantics of the emotional polarity is the key to affect recipients of the medical information. In this paper, we propose an adaptive learning emotion identification method (ALEIM) based on mutual information feature weight, which captures the correlation and redundancy of features. In order to evaluate the proposed method's effectiveness, we use four basic corpus libraries crawled from the Haodf's online platform and employ Taiwan University NTUSD Simplified Chinese Emotion Dictionary for emotion classification. The experimental results show that our proposed ALEIM method has a better performance for the identification of the low-frequency words' redundant features in comments of the online medical knowledge sharing community.

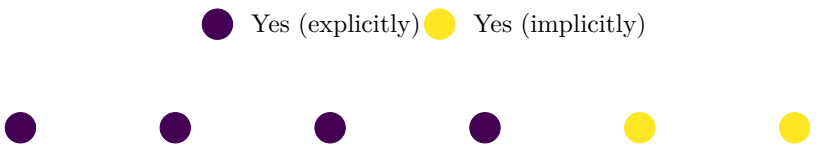


ID: 2858

Agreement Score: 0.33

Extracting and Clustering Main Ideas from Student Feedback Using Language Models

Feedback mechanisms for academic courses have been widely used to measure students **opinions** and satisfaction towards different components of a course; concurrently, open-text detailed impressions enable professors to continually improve their course. However, the process of reading through hundreds of student feedback responses across multiple subjects, followed by the extraction of important ideas is very time consuming. In this work, we propose an automated feedback summarizer to extract the main ideas expressed by all students on various components for each course, based on a pipeline integrating state-of-the-art Natural Language Processing techniques. Our method involves the usage of BERT language models to extract keywords for each course, identify relevant contexts for recurring keywords, and cluster similar contexts. We validate our tool on 8,201 feedback responses for 168 distinct courses from the Computer Science Department of University Politehnica of Bucharest for the 2019-2020 academic year. Our approach achieves a size reduction of 59% on the overall volume of text, while only increasing the mean average error when predicting course ratings from student open-text feedback by an absolute value of 0.06.

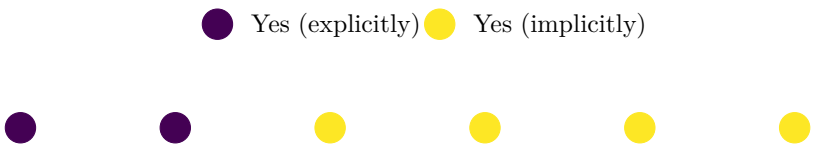


ID: 5400

Agreement Score: 0.33

Sampling techniques for Arabic Sentiment Classification: A comparative study

Over the last decade, the web 2.0 has been shifting the web to turn it into an **opinion** platform. This results in a large amount of raw data that overwhelms human capacity to extract valuable knowledge without assistance of machines. In real world applications, sentiment analysis faces imbalanced data problem. To tackle this problem, sampling techniques have been proposed. In this paper, we focus on studying the performance of these techniques on Imbalanced Data of Arabic Sentiment. We then conduct a comparative evaluation using Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF) as classification algorithms.

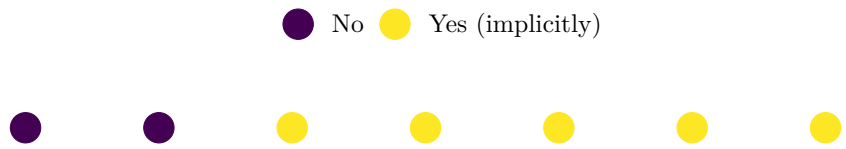


ID: 1696

Agreement Score: 0.33

Knowledge-Guided Sentiment Analysis Via Learning From Natural Language Explanations

Sentiment analysis is crucial for studying public **opinion** since it can provide us with valuable information. Existing sentiment analysis methods rely on finding the sentiment element from the content of user-generated. However, the question of why a message produces certain emotions has not been well explored or utilized in previous works. To address this challenge, we propose a natural language explanation framework for sentiment analysis that provides sufficient domain knowledge for generating additional labelled data for each new labelling decision. A rule-based semantic parser transforms these explanations into programmatic labelling functions that generate noisy labels for an arbitrary amount of unlabelled sentiment information to train a sentiment analysis classifier. Experiments on two sentiment analysis datasets demonstrate the superiority it achieves over baseline methods by leveraging explanations as external knowledge to joint training a sentiment analysis model rather than only labels. An ablation study is conducted to clarify the relative contribution of natural language explanations.

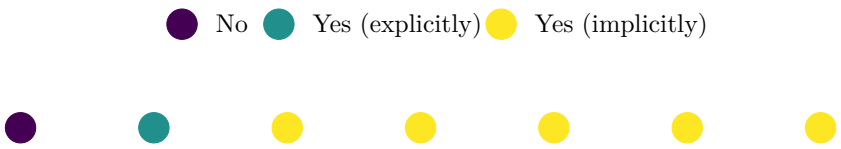


ID: 1749

Agreement Score: 0.43

Analyzing mass media influence using natural language processing and time series analysis

A key question of collective social behavior is related to the influence of mass media on public **opinion**. Different approaches have been developed to address quantitatively this issue, ranging from field experiments to mathematical models. In this work we propose a combination of tools involving natural language processing and time series analysis. We compare selected features of mass media news articles with measurable manifestation of public **opinion**. We apply our analysis to news articles belonging to the 2016 US presidential campaign. We compare variations in polls (as a proxy of public **opinion**) with changes in the connotation of the news (sentiment) or in the agenda (topics) of a selected group of media outlets. Our results suggest that the sentiment content by itself is not enough to understand the differences in polls, but the combination of topics coverage and sentiment content provides an useful insight of the context in which public **opinion** varies. The methodology employed in this work is far general and can be easily extended to other topics of interest.



ID: 1129

Agreement Score: 0.43

USING DATA FROM REDDIT, PUBLIC DELIBERATION, AND SURVEYS TO MEASURE PUBLIC OPINION ABOUT AUTONOMOUS VEHICLES

When and how can researchers synthesize survey data with analyses of social media content to study public **opinion**, and when and how can social media data complement surveys to better inform researchers and policymakers? This paper explores how public **opinions** might differ between survey and social media platforms in terms of content and audience, focusing on the test case of **opinions** about autonomous vehicles. The paper first extends previous overviews comparing surveys and social media as measurement tools to include a broader range of survey types, including surveys that result from public deliberation, considering the dialogic characteristics of different social media, and the range of issue publics and marginalized voices that different surveys and social media forums can attract. It then compares findings and implications from analyses of public **opinion** about autonomous vehicles from traditional surveys, results of public deliberation, and analyses of Reddit posts, applying a newly developed computational text analysis tool. Findings demonstrate that social media analyses can both help researchers learn more about issues that are uncovered by surveys and also uncover **opinions** from subpopulations with specialized knowledge and unique orientations toward a subject. In light of these findings, we point to future directions on how researchers and policymakers can synthesize survey and social media data, and the corresponding data integration techniques, to study public **opinion**.