Unleashing Context: Enhancing Tweet Classification with Contextual Insights

Andrija Banić, Marko Jurić, Dario Pavlović

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{andrija.banic,marko.juric,dario.pavlovic}@fer.hr

Abstract

In today's media-dominated world, the widespread flow of information increases the risk of encountering dangerous fake news. This alarming issue has attracted the attention of many researchers. However, there is a lack of labeled data available, and the existing labeled data is imbalanced, which poses a major challenge in supervised learning. With numerous research papers focusing on various neural language models and manually created features for classification, this study aims to investigate how the context of an individual tweet affects feature engineering and the interpretability of the results. Specifically, we focus on the task of stance classification, which involves determining the position or attitude expressed in a tweet towards a particular topic. In our investigation, we compare the performance of different feature engineering approaches when considering different levels of context. We explore three scenarios: (1) no context considered, (2) only the context of the tweet being replied to (referred to as „parent" context), and (3) both the replies and the tweet being replied to (referred to as „children" and „parent" context if we think of tweets represented as a tree structure). By examining the impact of context on classification accuracy and incorporating the concept of stance, we aim to understand the potential advantages of incorporating contextual information in tweet classification.

1. Introduction

Rumours have been a part of our lives since the day we were born, representing unofficial and intriguing stories or news that quickly circulate from person to person. In the modern era, these rumours have evolved from harmless gossip about friends and colleagues to potentially dangerous and concerning misinformation. With millions of tweets being posted daily, it is crucial to identify and distinguish those tweets that spread harmful rumours, as they are often convincing and challenging to detect manually. Therefore, it is essential to develop fast, precise, and automated methods to tackle this issue. Our paper concentrates on stance classification, which aims to determine the attitude of the author of a tweet towards a specific target in Twitter conversations. We can visualize these conversations as trees, where the root represents the source post initiating the topic, and the following replies form a thread that branches out. The main goal of stance classificaton is to classify the tweet into one of the four categories which are support, deny, query or comment (**SDQC**).

The objective of our task is to determine the stance towards an underlying rumour, which is not explicitly provided but can often be inferred from the source post or the root of the discussion thread. While several papers have addressed this challenge, as discussed in Section 2., our approach focuses on using simpler machine learning models, such as Random Forest Classifier (RF), Gradient Boosting Classifier (GBC), and Stochastic Gradient Descent Classifier (SGDC). Instead of relying on complex models like GPT, BERT or LSTM-based language models, our emphasis is on feature engineering, described in Subsection 4.2. and, more importantly, context shown in Section 5. By context, we mean incorporating the stance of children or parent tweets, as well as the word embedding difference between the tweet we are classifying and its children or parent tweet. Research has shown that the stance of replies can be highly indicative of the post being classified (Aggarwal and Aker, 2019). Hence, we conducted experiments considering three scenarios:

- No context: We classify the tweet using features extracted solely from the tweet text itself.
- Parent context: We classify the tweet using features extracted from the tweet text itself, along with the word embedding difference between the tweet and its parent, incorporating the stance of the parent as an additional feature.
- Parent and children context: Similar to the parent context, we include the children of the tweet (their stance and word embedding difference) being classified.

In this paper, we aim to present our findings and insights gained from these experiments. Our focus lies in leveraging contextual information and feature engineering rather than relying on more complex approaches. The following sections provide a detailed explanation of our methodology, experimental setup, and results, shedding light on the effectiveness of our approach in tackling the challenge of rumour stance classification.

2. Related Work

The rise of social media and social networking platforms, such as Facebook and Twitter, has led to an explosion of user-generated content, with millions of posts and tweets being shared daily. However, until the 2017 RumourEval competition, little attention was given to the veracity of these tweets. The competition aimed to determine the veracity and support for rumors, marking a turning point in the field (Derczynski et al., 2017). Two years later, another RumourEval competition took place, attracting a significantly larger number of participants. Although the results demonstrated improvements compared to the previous competition, with some teams achieving an F1 score as high as 0.6187 in the task of rumour stance classification, the top three teams employed different techniques and models, approaching the problem from multiple angles (Gorrell et al., 2019). Of particular interest is the team that achieved the best performance in stance classification by utilizing a model that has recently revolutionized our lives - the GPT (Generative Pre-trained Transformer) model. This approach involved incorporating feature extraction with the GPT model (Yang et al., 2019). It is worth noting that the most common approach among participants was to create hand-crafted features that would effectively capture the semantic meaning of the tweet (Aker et al., 2017). Conversely, the second-ranked team decided to take a different approach, employing the BERT (Bidirectional Encoder Representations from Transformers) architecture. BERT was a groundbreaking development at the time, as it was introduced in 2018, following the introduction of transformers in 2017 (Devlin et al., 2018). Remarkably, recent studies have revealed that pre-trained language models surpass the performance of LSTM (Long Short-Term Memory)-based language models, which had long been considered the gold standard for stance classification (B. Bharathi and Balaji, 2020).

Expanding our investigation to sentiment analysis, we encounter another difficulty. As indicated by the title of a recent study by (Kenyon-Dean et al., 2018) implies, sentiment analysis is a complex and challenging task in its own right. However, it has been demonstrated that sentiment features can be useful for stance classification, although they are not sufficient on their own (Sobhani et al., 2016). A key distinguishing factor between sentiment analysis and stance classification lies in the consideration of the text's surrounding context. In sentiment analysis, the focus is primarily on analyzing the sentiment or emotional tone expressed within an individual text, without incorporating additional contextual information. This approach seeks to determine whether the sentiment conveyed is positive, negative, or neutral. On the other hand, stance classification involves determining the position or attitude expressed in a text towards a specific topic. Unlike sentiment analysis, stance classification takes into account the surrounding tweets or the broader context in which the text is situated. By considering the tweets preceding or following a particular text, we can gain insights into the stance of the author, such as whether they support, oppose, or are neutral towards the discussed topic.

In our paper, we explore the use of a Polarity Sentiment Analysis feature, which assesses the sentiment polarity of individual tweets, for enhancing stance classification. However, we acknowledge that relying solely on sentiment features is not sufficient for accurate stance classification. Interestingly, in the RumourEval competition, it was observed that the majority of models submitted primarily relied on the stance of the replies while neglecting the parent tweet as a source of context. This paper aims to shed light on the distinction between using no context at all for tweet stance classification, employing only the context of the „parents“ , and finally, using both the „children“ and „parent“ context. By exploring the role of context in determining classification accuracy, we aim to better understand the factors that contribute to effectively categorizing tweets. Our objective is to gain insights into how considering the surrounding information can enhance the accuracy of tweet classification.

3. Dataset Analysis

In our study, we utilized the RumourEval2019 dataset as the primary data source. This dataset offered a valuable resource for our research objectives. It comprised a collection of tweets and Reddit posts, with the Reddit posts modified to match the format of tweet posts. Although the Reddit data was included, it was not used in this paper. The inclusion of both Twitter and Reddit data was particularly advantageous due to the substantial user base, extensive posting activity, and the wealth of information available on these platforms. The dataset was curated by gathering posts from Twitter during various crisis events, including occurrences such as natural disasters or public emergencies. These crisis events serve as topics within the dataset, and each topic consists of a source tweet accompanied by a set of reply tweets. The organization of these tweets follows a tree-like structure, which captures the conversational relationships and hierarchy between the source tweet and its corresponding replies. To ensure the comprehensiveness of the dataset, the tweets were labeled using crowdsourcing platforms. These labels assign one of four categories to each tweet: *supporting*, *denying*, *query*, or *commenting*. This labeling process provides valuable ground truth information that enables the training and evaluation of models for rumor detection and veracity prediction.

- Support: Tweets categorized as „Support“ indicate that the author expresses agreement, endorsement, or validation towards the target under discussion. They share a favorable stance and indicate support for the viewpoint related to the topic.

- Deny: Tweets classified as „Deny“ reflect a contradictory stance or rejection of the target. Authors in this category express disagreement, disbelief, or negation towards the viewpoint associated with the topic of discussion.

- Query: Tweets falling into the „Query“ category signify curiosity, doubt, or a request for further information or clarification about the target. Authors seek more details, explanations, or evidence related to the topic, indicating a neutral or uncertain stance.

- Comment: The „Comment“ category encompasses tweets that do not clearly align with the previously mentioned categories. Tweets in this category serve as general remarks, observations, or opinions on the topic without explicitly expressing support, denial, or inquiry. They might provide additional context, personal experiences, or commentary related to the target.

As depicted in Figure 1, a notable characteristic of the dataset is the significant label imbalance present among the different categories. This label imbalance introduces a challenge due to the disproportionate representation of the

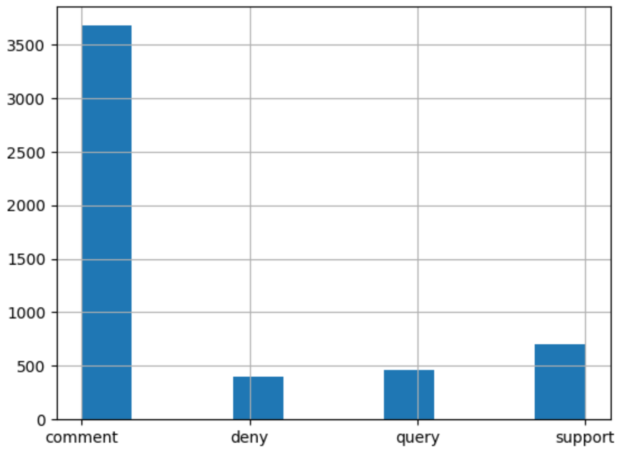

Figure 1: The distribution of labels for the source tweet reveals a significant disparity, with a substantial majority of data being labeled as „comment“. This disproportionate distribution poses a challenge as it can lead to models becoming biased towards predicting the majority label.

comment label, which greatly outnumbers the other labels. Consequently, this label imbalance can potentially lead to biases in the prediction of tweets, favoring the classification of tweets as *comments* due to their prevalence in the dataset. The dataset itself is also relatively small in size, which further impacts label imbalance. With a limited number of instances available for each label category, the model's ability to learn and generalize effectively can be compromised, particularly for the minority classes.

4. Experimental Setup

The setup for the experiment consists of data selection, preprocessing, feature extraction and choosing the optimal models for the analysis. By selecting relevant data and preprocessing the text, the aim is to extract meaningful information and remove unnecessary noise or artifacts that might interfere with the subsequent analysis. These steps help to standardize the text data, make it more manageable, and enable more effective analysis and modeling.

4.1. Data selection and preprocessing

Given the vast amount of data and metadata associated with each tweet, it is necessary to carefully select the relevant information for analysis. In our study, we specifically consider the following data: *favourite count*, *text*, *hashtags*, *retweet count*, and a list of *direct replies*. These selected features are deemed essential for capturing key aspects of the tweet content, engagement, and contextual information. To prepare the text data for embedding feature, we conducted a preprocessing step. This step involved several procedures aimed at enhancing the quality and consistency of the text. Specifically, we performed lemmatization to transform words into their base or dictionary form. Furthermore, we implemented a cleaning process to remove irrelevant elements such as external links, punctuation marks, stop words, mentions of other users, and symbols. Additionally, we replaced numeric values within the text with the hashtag sign as these numeric values were deemed non-informative for our analysis.

It is important to note that while we modified certain aspects of the text during preprocessing, the original text remains intact. This allows us to create additional features based on the original text, such as determining whether the tweet has a link, detecting mentions of news agencies, or quantifying the amount of question and exclamation marks present in the text. These additional features provide valuable insights and contribute to a more comprehensive analysis of the tweet content.

4.2. Features

The extracted features are the following:

- Capital Ratio: This feature calculates the ratio of uppercase characters in the entire text. It can provide insights into the text's writing style or emphasis. For example, a higher capital ratio might indicate a more assertive or emphatic tone, which could be relevant in determining the stance of the reply.

- Length of the Text: The length of the text refers to the number of characters or words in the text. Longer texts might suggest that the reply is attempting to provide more detailed explanations or arguments, potentially indicating a denying stance. Shorter texts, on the other hand, could imply a supporting or commenting stance that doesn't require extensive explanations.

- Polarity Sentiment Analysis: In our study, we utilize the TextBlob library to perform polarity sentiment analysis. This analysis involves assigning a polarity score to the text, which indicates the sentiment expressed in the tweet or reply. The polarity score ranges from -1 to 1, where -1 represents a negative sentiment, 0 indicates a neutral sentiment, and 1 reflects a positive sentiment. For example, tweets with a positive polarity score may indicate support for the target, while those with a negative polarity score might suggest denial or disagreement.

- Amount of Question and Exclamation Marks: This feature counts the number of question marks and exclamation marks in the text. The presence of these punctuation marks can convey different intentions or emotional tones. For example, a higher count of question marks might indicate a questioning or uncertain stance, while exclamation marks could imply strong emotions or emphasis.

- Word Embeddings: They capture the semantic meaning of words by representing them as dense vectors in a high-dimensional space. In this case, the Gensim library is used with pre-trained glove vectors *glove-twitter-25* to obtain word embeddings for the text. These embeddings encode contextual information and can help capture the meaning and relationships between words in the text.

- Has Link: This feature checks whether the post contains a link. The presence of a link might suggest that

Table 1: Accuracy and F1-scores of different models without context, with parent context and with both the parent and the reply context

Model	Macro F1-score	Accuracy	$F1_S$	$F1_D$	$F1_Q$	$F1_C$
Without context						
Random Forest Classifier	0.2624	0.76	0.00	0.00	0.18	0.87
Gradient Boosting Classifier	0.2953	0.71	0.14	0.00	0.21	0.84
Stohastic Gradient Descent Classifier	**0.3735**	0.72	**0.23**	0.00	**0.42**	0.84
With parent context						
Random Forest Classifier	0.2971	0.77	0.00	0.00	0.32	0.87
Gradient Boosting Classifier	**0.3454**	0.69	0.15	0.07	0.35	0.82
Stohastic Gradient Descent Classifier	0.2925	0.72	0.14	0.07	0.41	0.81
With children and parent context						
Random Forest Classifier	0.2964	0.77	0.02	0.00	0.29	0.87
Gradient Boosting Classifier	0.3062	0.69	0.14	0.04	0.23	0.82
Stohastic Gradient Descent Classifier	**0.3641**	0.72	0.21	**0.18**	0.36	0.71

the reply includes a source or reference, potentially indicating a supporting or denying stance by providing external evidence.

- Mentions News Agency: This feature detects whether the post mentions a news agency. This can be relevant in determining the credibility or reliability of the information shared. If a reply mentions a news agency, it might indicate a supporting or denying stance based on the perceived trustworthiness of the news source.

5. Results

Our primary objective was to investigate the impact of context on the performance of different models using three different scenarios: without context, with context from the parent tweet only, and with context from both parent and child tweets. The experimental results are presented in Table 1, where each model is evaluated under the previously mentioned scenarios. To determine the optimal model, we employed the Grid Search algorithm to search for the best hyperparameters. The results indicate that the Stochastic Gradient Descent Classifier (SGDC) outperforms the other models in terms of the overall Macro F1-score, except for when we only consider the parent context. An analysis of the results suggests that the Gradient Boosting Classifier (GBC) outperforms the Stochastic Gradient Descent Classifier (SGDC) which we can conclude on several observations, including the macro F1 score and the F1 scores of other labels, which predominantly favor the GBC model. Notably, when the SGDC model is trained without any context from other tweets, it achieves the best F1-scores for most labels, except for the *deny* label. The *supporting* and *deny* label are particularly significant as they indicate the potential presence of a rumor in the tweet they respond to.

Considering the importance of accurately identifying the *supporting* and *deny* labels, the optimal model for classifying our dataset would be the SGDC model with children and parent context. Although the F1-scores for other labels are not significantly different from the alternative scenarios, the F1-score for the *deny* label improves considerably compared to other models. Thus, based on the evaluation results, we recommend utilizing the SGDC model with context from both parent and child tweets as it offers competitive performance across all labels.

6. Conclusion

In recent years, there has been a significant effort to tackle the challenging problem of rumour veracity. However, this task has proven to be exceptionally difficult due to various factors, including the limited availability of labeled datasets. These datasets are often small in size and untrustworthy, making them inadequate for training accurate models. Another challenge in determining rumour veracity lies in classifying the stance of text, which has also been shown to be problematic due to imbalanced label distributions.

This paper specifically focuses on the stance classification problem using simple machine learning models. Instead of relying on complex models, we employ straightforward machine learning algorithms and explore the effectiveness of different hand-crafted features. However, our primary emphasis lies in investigating the impact of context on stance classification.

To substantiate our claim that context plays a crucial role in classifying stances accurately, it is essential to conduct experiments using more sophisticated models on larger datasets. By doing so, we can further validate the significance of context in enhancing the performance of stance classification models.

References

Piush Aggarwal and Ahmet Aker. 2019. Identification of good and bad news on Twitter. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 9–17, Varna, Bulgaria, September. INCOMA Ltd.

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*

2017, pages 31–39, Varna, Bulgaria, September. INCOMA Ltd.

J. Bhuvana B. Bharathi and Nitin Nikamanth Appiah Balaji. 2020. Textual and contextual stance detection from tweets using machine learning approach.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada, August. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. volume abs/1810.04805.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana, June. Association for Computational Linguistics.

Parinaz Sobhani, Saif Mohammad, and Svetlana Kiritchenko. 2016. Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 159–169, Berlin, Germany, August. Association for Computational Linguistics.

Ruoyao Yang, Wanying Xie, Chunhua Liu, and Dong Yu. 2019. BLCU_NLP at SemEval-2019 task 7: An inference chain-based GPT model for rumour evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1090–1096, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.