

# Stance Classification for Rumour Analysis in Twitter: Exploiting Affective Information and Conversation Structure

Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti  
Dipartimento di Informatica  
Università degli Studi di Torino  
{pamungka, basile, patti}@di.unito.it

## Abstract

Analysing how people react to rumours associated with news in social media is an important task to prevent the spreading of misinformation, which is nowadays widely recognized as a dangerous tendency. In social media conversations, users show different stances and attitudes towards rumourous stories. Some users take a definite stance, supporting or denying the rumour at issue, while others just comment it, or ask for additional evidence on the rumour's veracity. A shared task has been proposed at SemEval-2017 (Task 8, Sub-Task A), which is focused on rumour stance classification in English tweets. The goal is predicting user stance towards emerging rumours in Twitter, in terms of supporting, denying, querying, or commenting the original rumour, looking at the conversation threads originated by the rumour. This paper describes a new approach to this task, where the use of conversation-based and affective-based features, covering different facets of affect, is explored. Our classification model outperforms the best-performing systems for stance classification at SemEval-2017 showing the effectiveness of the feature set proposed.

Copyright © CIKM 2018 for the individual papers by the papers' authors. Copyright © CIKM 2018 for the volume as a collection by its editors. This volume and its papers are published under the Creative Commons License Attribution 4.0 International (CC BY 4.0).

## 1 Introduction

Nowadays, people increasingly tend to use social media like Facebook and Twitter as their primary source of information and news consumption. There are several reasons behind this tendency, such as the simplicity to gather and share the news and the possibility of staying abreast of the latest news and updated faster than with traditional media. An important factor is also that people can be engaged in conversations on the latest breaking news with their contacts by using these platforms. Pew Research Center's newest report<sup>1</sup> shows that two-thirds of U.S. adults gather their news from social media, where Twitter is the most used platform. However, the absence of a systematic approach to do some form of fact and veracity checking may also encourage the spread of rumourous stories and misinformation [PVV13]. Indeed, in social media, unverified information can spread very quickly and becomes viral easily, enabling the diffusion of false rumours and fake information.

Within this scenario, it is crucial to analyse people attitudes towards rumours in social media and to resolve their veracity as soon as possible. Several approaches have been proposed to check the rumour veracity in social media [SSW<sup>+</sup>17]. This paper focus on a stance-based analysis of event-related rumours, following the approach proposed at SemEval-2017 in the new RumourEval shared task (Task 8, sub-task A) [DBL<sup>+</sup>17]. In this task English tweets from conversation threads, each associated to a newsworthy event and the rumours around it, are provided as data. The goal is to determine whether a tweet in the thread is supporting, denying, querying, or commenting the original rumour which started the conversation. It can be considered a stance classification task, where we

<sup>1</sup><http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/>

have to predict the user’s stance towards the rumour from a tweet, in the context of a given thread. This task has been defined as *open stance classification task* and is conceived as a key step in rumour resolution, by providing an analysis of people reactions towards an emerging rumour [PVV13, ZLP<sup>+</sup>16]. The task is also different from detecting stance towards a specific target entity [MKS<sup>+</sup>16].

**Contribution** We describe a novel classification approach, by proposing a new feature matrix, which includes two new groups: (a) features exploiting the conversational structure of the dataset [DBL<sup>+</sup>17]; (b) affective features relying on the use of a wide range of affective resources capturing different facets of sentiment and other affect-related phenomena. We were also inspired by the fake news study on Twitter in [VRA18], showing that false stories inspire fear, disgust, and surprise in replies, while true stories inspire anticipation, sadness, joy, and trust. Meanwhile, from a dialogue act perspective, the study of [NS13] found that a relationship exists between the use of an affective lexicon and the communicative intention of an utterance which includes AGREE-ACCEPT (support), REJECT (deny), INFO-REQUEST (question), and OPINION (comment). They exploited several LIWC categories to analyse the role of affective content.

Our results show that our model outperforms the state of the art on the Semeval-2017 benchmark dataset. Feature analysis highlights the contribution of the different feature groups, and error analysis is shedding some light on the main difficulties and challenges which still need to be addressed.

**Outline** The paper is organized as follows. Section 2 introduces the SemEval-2017 Task 8. Section 3 describes our approach to deal with open stance classification by exploiting different groups of features. Section 4 describes the evaluation and includes a qualitative error analysis. Finally, Section 5 concludes the paper and points to future directions.

## 2 SemEval-2017 Task 8: RumourEval

The SemEval-2017 Task 8 Task A [DBL<sup>+</sup>17] has as its main objective to determine the stance of the users in a Twitter thread towards a given rumour, in terms of support, denying, querying or commenting (SDQC) on the original rumour. Rumour is defined as a “*circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety so as to motivate finding out the actual truth*” [ZLP<sup>+</sup>15]. The task was very timing due to the growing importance of rumour resolution in the breaking news and to the urgency of preventing the spreading of misinformation.

Development Data				
Rumour	S	D	Q	C
Germanwings	69	11	28	173
Training Data				
Rumour	S	D	Q	C
Charlie Hebdo	239	58	53	721
Ebola Essien	6	6	1	21
Ferguson	176	91	99	718
Ottawa Shooting	161	76	63	477
Prince Toronto	21	7	11	64
Putin Missing	18	6	5	33
Sydney Siege	220	89	98	700
<b>Total</b>	<b>841</b>	<b>333</b>	<b>330</b>	<b>2734</b>
Testing Data				
Rumour	S	D	Q	C
Ferguson	15	4	17	66
Ottawa Shooting	10	2	20	91
Sydney Siege	5	1	12	69
Charlie Hebdo	9	2	8	74
Germanwings	11	5	15	71
Marina Joyce	5	30	10	110
Hillary’s Illness	39	27	24	297
<b>Total</b>	<b>94</b>	<b>71</b>	<b>106</b>	<b>778</b>

Table 1: Semeval-2017 Task 8 (A) dataset distribution.

**Dataset**<sup>2</sup> The data for this task are taken from Twitter conversations about news-related rumours collected by [ZLP<sup>+</sup>16]. They were annotated using four labels (SDQC): *support* - *S* (when tweet’s author support the rumour veracity); *deny* - *D* (when tweet’s author denies the rumour veracity); *query* - *Q* (when tweet’s author ask for additional information/evidence); *comment* - *C* (when tweet’s author just make a comment and does not give important information to asses the rumour veracity). The distribution consists of three sets: development, training and test sets, as summarized in Table 1, where you can see also the label distribution and the news related to the rumors discussed. Training data consist of 297 Twitter conversations and 4,238 tweets in total with related direct and nested replies, where conversations are associated to seven different breaking news. Test data consist of 1049 tweets, where two new rumourous topics were added.

**Participants** Eight teams participated in the task. The best performing system was developed by Turing (78.4 in accuracy). ECNU, MamaEdha, UWaterloo, and DFKI-DKT utilized ensemble classifier. Some systems also used deep learning techniques, including Turing, IKM, and MamaEdha. Meanwhile, NileTRMG and IITP used classical classifier (SVM) to

<sup>2</sup><http://alt.qcri.org/semeval2017/task8/index.php?id=data-and-tools>

build their systems. Most of the participants exploited word embedding to construct their feature space, beside the Twitter domain features.

### 3 Proposed Method

We developed a new model by exploiting several stylistic and structural features characterizing Twitter language. In addition, we propose to utilize conversational-based features by exploiting the peculiar tree structure of the dataset. We also explored the use of affective based feature by extracting information from several affective resources including dialogue-act inspired features.

#### 3.1 Structural Features

They were designed taking into account several Twitter data characteristics, and then selecting the most relevant features to improve the classification performance. The set of structural features that we used is listed below.

**Retweet Count:** The number of retweet of each tweet.

**Question Mark:** presence of question mark ”?”; binary value (0 and 1).

**Question Mark Count:** number of question marks present in the tweet.

**Hashtag Presence:** this feature has a binary value 0 (if there is no hashtag in the tweet) or 1 (if there is at least one hashtag in the tweet).

**Text Length:** number of characters after removing Twitter markers such as hashtags, mentions, and URLs.

**URL Count:** number of URL links in the tweet.

#### 3.2 Conversation Based Features

These features are devoted to exploit the peculiar characteristics of the dataset, which have a tree structure reflecting the conversation thread<sup>3</sup>.

**Text Similarity to Source Tweet:** Jaccard Similarity of each tweet with its source tweet.

**Text Similarity to Replied Tweet:** the degree of similarity between the tweet with the previous tweet in the thread (the tweet is a reply to that tweet).

**Tweet Depth:** the depth value is obtained by counting the node from sources (roots) to each tweet in their hierarchy.

<sup>3</sup>The implementation of these features is inspired from unpublished shared code [Gra17].

#### 3.3 Affective Based Features

The idea to use affective features in the context of our task was inspired by recent works on fake news detection, focusing on emotional responses to true and false rumors [VRA18], and by the work in [NS13] reflecting on the role of affect in dialogue acts [NS13]. Multifaceted affective features have been already proven to be effective in some related tasks [LFPR16], including the stance detection task proposed at SemEval-2016 (Task 6).

We used the following affective resources relying on different emotion models.

**Emolex:** it contains 14,182 words associated with eight primary emotion based on the Plutchik model [MT13, Plu01].

**EmoSenticNet (EmoSN):** it is an enriched version of SenticNet [COR14] including 13,189 words labeled by six Ekman’s basic emotion [PGH<sup>+</sup>13, Ekm92].

**Dictionary of Affect in Language (DAL):** includes 8,742 English words labeled by three scores representing three dimensions: Pleasantness, Activation and Imagery [Whi09].

**Affective Norms for English Words (ANEW):** consists of 1,034 English words [BL99] rated with ratings based on the Valence-Arousal-Dominance (VAD) model [OST57].

**Linguistic Inquiry and Word Count (LIWC):** this psycholinguistic resource [PFB01] includes 4,500 words distributed into 64 emotional categories including positive (PosEMO) and negative (NegEMO).

#### 3.4 Dialogue-Act Features

We also included additional 11 categories from LIWC, which were already proven to be effective in dialogue-act task in previous work [NS13]. Basically, these features are part of the affective feature group, but we present them separately because we are interested in exploring the contribution of such feature set separately. This feature set was obtained by selecting 4 communicative goals related to our classes in the stance task: **agree-accept** (support), **reject** (deny), **info-request** (question), and **opinion** (comment). The 11 LIWC categories include:

**Agree-accept:** Assent, Certain, Affect;

**Reject:** Negate, Inhib;

**Info-request:** You, Cause;

**Opinion:** Future, Sad, Insight, Cogmech.

No.	Systems	Accuracy
1.	Turing’s System	78.4
2.	Aker et al. System	79.02
3.	Our System	<b>79.5</b>
	RumourEval Baseline	74.1

Table 2: Results and comparison with state of the art

## 4 Experiments, Evaluation and Analysis

We used the RumourEval dataset from SemEval-2017 Task 8 described in Section 2. We defined the rumour stance detection problem as a simple four-way classification task, where every tweet in the dataset (source and direct or nested reply) should be classified into one among four classes: support, deny, query, and comment. We conducted a set of experiments in order to evaluate and analyze the effectiveness of our proposed feature set.<sup>4</sup>

The results are summarized in Table 2, showing that our system outperforms all of the other systems in terms of accuracy. Our best result was obtained by a simple configuration with a support vector classifier with radial basis function (RBF) kernel. Our model performed better than the best-performing systems in SemEval 2017 Task 8 Subtask A (Turing team, [KLA17]), which exploited deep learning approach by using LTSM-Branch model. In addition, we also got a higher accuracy than the system described in [ADB17], which exploits a Random Forest classifier and word embeddings based features.

We experimented with several classifiers, including Naive Bayes, Decision Trees, Support Vector Machine, and Random Forest, noting that SVM outperforms the other classifiers on this task. We explored the parameter space by tuning the SVM hyperparameters, namely the penalty parameter  $C$ , kernel type, and class weights (to deal with class imbalance). We tested several values for  $C$  (0.001, 0.01, 0.1, 1, 10, 100, and 1000), four different kernels (linear, RBF, polynomial, and sigmoid) and weighted the classes based on their distribution in the training data. The best result was obtained with  $C=1$ , RBF kernel, and without class weighting.

An ablation test was conducted to explore the contribution of each feature set. Table 5 shows the result of our ablation test, by exploiting several feature sets on the same classifier (SVM with RBF kernel)<sup>5</sup>. This evaluation includes macro-averages of precision, recall and  $F_1$ -score as well as accuracy. We also presented

<sup>4</sup>We built our system by using scikit-learn Python Library: <http://scikit-learn.org/>

<sup>5</sup>Source code is available on the GitHub platform: <https://github.com/dadangewp/SemEval2017-RumourEval>

	S	D	Q	C
<b>Support</b>	<b>27</b>	0	3	64
<b>Deny</b>	2	<b>0</b>	1	68
<b>Query</b>	0	0	<b>50</b>	56
<b>Comment</b>	13	0	8	<b>757</b>

Table 3: Confusion Matrix

	S	D	Q	C
<b>Support</b>	<b>39</b>	14	5	13
<b>Deny</b>	8	<b>28</b>	5	30
<b>Query</b>	2	3	<b>62</b>	4
<b>Comment</b>	14	14	2	<b>41</b>

Table 4: Confusion Matrix on Balanced Dataset

the scores for each class in order to get a better understanding of our classifier’s performance.

Using only conversational, affective, or dialogue-act features (without structural features) did not give a good classification result. Set B (conversational features only) was not able to detect the query and deny classes, while set C (affective features only) and D (dialogue-act features only) failed to catch the support, query, and deny classes. Conversational features were able to improve the classifier performance significantly, especially in detecting the support class. Sets E, H, I, and K which utilize conversational features induce an improvement on the prediction of the support class (roughly from 0.3 to 0.73 on precision). Meanwhile, the combination of affective and dialogue-act features was able to slightly improve the classification of the query class. The improvement can be seen from set E to set K where the  $F_1$ -score of query class increased from 0.52 to 0.58. Overall, the best result was obtained by the K set which encompasses all sets of features. It is worth to be noted that in our best configuration system, not all of affective and dialogue-act features were used in our feature vector. After several optimization steps, we found that some features were not improving the system’s performance. Our final list of affective and dialogue-act based features includes: **DAL Activation**, **ANEW Dominance**, **Emolex Negative**, **Emolex Fear**, **LIWC Assent**, **LIWC Cause**, **LIWC Certain** and **LIWC Sad**. Therefore, we have only 17 columns of features in the best performing system covering structural, conversational, affective and dialogue-act features.

We conducted a further analysis of the classification result obtained by the best performing system (79.50 on accuracy). Table 3 shows the confusion matrix of our result. On the one hand, the system is able to detect the comment tweets very well. However, this result is biased due to the number of comment data in

Ablation Test		Overall				Support				Query				Comment			
Set	Features	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
A	Structural	0.731	0.41	0.37	0.38	0.18	<b>0.28</b>	0.18	0.22	0.39	0.56	0.39	0.46	0.91	0.78	0.91	0.84
B	Conversational	0.767	0.42	0.31	0.33	0.29	0.93	0.29	0.44	0	0	0	0	1	0.76	1	0.87
C	Affective	0.742	0.19	0.25	0.21	0	0	0	0	0	0	0	0	1	0.74	1	0.85
D	Dialogue-Act	0.742	0.19	0.25	0.21	0	0	0	0	0	0	0	0	1	0.74	1	0.85
E	A + B	0.783	0.54	0.43	0.45	0.29	<b>0.73</b>	0.29	0.41	0.42	0.62	0.42	<b>0.52</b>	0.96	0.8	0.96	0.87
F	A + C	0.741	0.42	0.36	0.38	0.14	0.27	0.14	0.18	0.39	0.62	0.39	0.48	0.93	0.77	0.93	0.84
G	A + D	0.736	0.42	0.37	0.38	0.18	0.3	0.18	0.23	0.37	0.59	0.37	0.45	0.92	0.77	0.92	0.84
H	E + C	0.788	0.56	0.42	0.46	0.28	<b>0.74</b>	0.28	0.4	0.44	0.7	0.44	0.54	0.97	0.8	0.97	0.87
I	E + D	0.784	0.53	0.43	0.46	0.3	<b>0.65</b>	0.3	0.41	0.45	0.67	0.45	0.54	0.96	0.8	0.96	0.87
J	F + D	0.749	0.43	0.36	0.38	0.14	0.33	0.14	0.19	0.38	0.63	0.38	0.47	0.94	0.77	0.94	0.85
K	All Features	<b>0.795</b>	0.57	0.43	0.47	0.29	<b>0.73</b>	0.29	0.41	0.47	0.75	0.47	<b>0.58</b>	0.97	0.8	0.97	0.88

\*deny class is not presented, since the score is always zero (0)

Table 5: Ablation test on several feature sets.

the dataset. On the other hand, the system is failing to detect denying tweets, which were falsely classified into comments (68 out of 71)<sup>6</sup>. Meanwhile, approximately two thirds of supporting tweets and almost half of querying tweets were classified into the correct class by the system.

In order to assess the impact of class imbalance on the learning, we performed an additional experiment with a balanced dataset using the best performing configuration. We took a subset of the instances equally distributed with respect to their class from the training set (330 instances for each class) and test set (71 instances for each class). As shown in Table 4, our classifier was able to correctly predict the underrepresented classes much better, although the overall accuracy is lower (59.9%). The result of this analysis clearly indicates that class imbalance has a negative impact on the system performance.

#### 4.1 Error analysis

We conducted a qualitative error analysis on the 215 misclassified in the test set, to shed some light on the issues and difficulties to be addressed in future work and to detect some notable error classes.

**Denying by attacking the rumour’s author.** An interesting finding from the analysis of the Marina Joyce rumour data is that it contains a lot of denying tweets including insulting comments towards the author of the source tweet, like in the following cases:

*Rumour:* Marina Joyce

*Misclassified tweets:*

(da1) stfu you toxic sludge

(da2) @sampepper u need rehab 🙄

*Misclassification type:* deny (gold)  $\rightsquigarrow$  comment (prediction)

*Source tweet:*

(s1) Anyone who knows Marina Joyce personally knows she has a serious drug addic-

tion. she needs help, but in the form of rehab  
#savemarinajoyce

Tweets like (da1) and (da2) seem to be more inclined to show the respondent’s personal hatred towards the s1-tweet’s author than to deny the veracity of the rumour. In other words, they represent a peculiar form of denying the rumour, which is expressed by personal attack and by showing negative attitudes or hatred towards the rumour’s author. This is different from denying by attacking the source tweet content, and it was difficult to comprehend for our system, that often misclassified such kind of tweets as comments.

**Noisy text, specific jargon, very short text.** In (da1) and (da2) (as in many tweets in the test set), we also observe the use of noisy text (abbreviations, misspellings, slang words and slurs, question statements without question mark, and so on) that our classifier struggles to handle. Moreover, especially in tweets from the Marina Joyce rumour’s group, we found some very short tweets in the denying class that do not provide enough information, e.g. tweets like “shut up!”, “delete”, and “stop it. get some help”.

**Argumentation context.** We also observed misclassification cases that seem to be related to a deeper capability of dealing with the argumentation context underlying the conversation thread.

*Rumour:* Ferguson

*Misclassified tweet:*

(arg1)@QuadCityPat @AP I join you in this demand. Unconscionable.

*Misclassification type:* deny (gold)  $\rightsquigarrow$  comment (prediction)

*Source tweet:*

(s2) @AP I demand you retract the lie that people in #Ferguson were shouting “kill the police”, local reporting has refuted your ugly racism

Here the misclassified tweet is a reply including an explicit expression of agreement with the author of the

<sup>6</sup>A similar observation is reported by the best team at Semeval-2017 [KLA17].

source tweet (“*I join you*”). Tweet (s2) is one of the rare cases of source tweets denying the rumor (source tweets in the RumourEval17 dataset are mostly supporting the rumor at issue). Our hypothesis is that it is difficult for a system to detect such kind of stance without a deeper comprehension of the argumentation context (e.g., if the author’s stance is denying the rumor, and I agree with him, then I am denying the rumor as well). In general, we observed that when the source tweet is annotated by the *deny* label, most of denying replies of the thread include features typical of the support class (and vice versa), and this was a criticism.

**Mixed cases.** Furthermore, we found some borderline mixed cases in the gold standard annotation. See for instance the following case:

*Rumour:* Ferguson

*Misclassified tweet:*

(mx1) @MichaelSkolnik @MediaLizzy Oh do tell where they keep track of ”vigilante” stats. That’s interesting.

*Misclassification type:* query (gold)  $\rightsquigarrow$  comment (prediction)

*Source tweet:*

(s3) Every 28 hours a black male is killed in the United States by police or vigilantes. #Ferguson

Tweet (mx1) is annotated with a *query* label rather than as a *comment* (our system prediction), but we can observe the presence of a comment (“*That’s interesting*”) after the request for clarification, so it seems to be a kind of mixed case, where both labels make sense.

**Citation of the source’s tweet.** We have noticed many misclassified cases of replying tweets with error pattern *support* (gold)  $\rightsquigarrow$  *comment* (our prediction), where the text contains a literal citation of the source tweet, like in the following tweet: *THIS HAS TO END “@MichaelSkolnik: Every 28 hours a black male is killed in the United States by police or vigilantes. #Ferguson”* (the text enclosed in quotes is the source tweet). Such kind of mistakes could be maybe addressed by applying some pre-processing to the data, for instance by detecting the literal citation and replacing it with a marker.

**Figurative language devices.** Finally, the use of figurative language (e.g., sarcasm) is also an issue that should be considered for the future work. Let us consider for instance the following misclassified tweets:

*Rumour:* Hillary’s Illness

*Misclassified tweets:*

(fg1) @mitchellvii True, after all she can open

a pickle jar.

(fg2) @mitchellvii Also, except for having a 24/7 MD by her side giving her Valium injections, Hillary is in good health! <https://t.co/GieNxxTXX7>

(fg3) @mitchellvii @JoanieChesnutt At the very peak yes, almost time to go down a cliff and into the earth.

*Misclassification type:* support (gold)  $\rightsquigarrow$  comment (prediction)

*Source tweet:*

(s4) Except for the coughing, fainting, apparent seizures and ”short-circuits,” Hillary is in the peak of health.

All misclassified tweets (fg1-fg3) from the *Hillary’s illness* data are replies to a source tweet (s4), which is featured by sarcasm. In such replies authors support the rumor by echoing the sarcastic tone of the source tweet. Such more sophisticated cases, where the supportive attitude is expressed in an implicit way, were challenging for our classifier, and they were quite systematically misclassified as simple comments.

## 5 Conclusion

In this paper we proposed a new classification model for rumour stance classification. We designed a set of features including structural, conversation-based, affective and dialogue-act based feature. Experiments on the SemEval-2017 Task 8 Subtask A dataset show that our system based on a limited set of well-engineered features outperforms the state-of-the-art systems in this task, without relying on the use of sophisticated deep learning approaches. Although achieving a very good result, several research challenges related to this task are left open. Class imbalance was recognized as one the main issues in this task. For instance, our system was struggling to detect the deny class in the original dataset distribution, but it performed much better in that respect when we balanced the distribution across the classes.

A re-run of the RumourEval shared task has been proposed at SemEval 2019<sup>7</sup> and it will be very interesting to participate to the new task with an evolution of the system here described.

## Acknowledgements

Endang Wahyu Pamungkas, Valerio Basile and Viviana Patti were partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618\_L2\_BOSC\_01).

<sup>7</sup><http://alt.qcri.org/semEval2019/>

## References

- [ADB17] Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. Simple open stance classification for rumour analysis. In *Proc. of RANLP 2017*, pages 31–39. INCOMA Ltd., 2017.
- [BL99] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical Report C-1, The Center for Research in Psychophysiology, University of Florida., 1999.
- [COR14] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proc. of AAAI 2014*, 2014.
- [DBL<sup>+</sup>17] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureal: Determining rumour veracity and support for rumours. In *Proc. of SemEval-2017*, pages 69–76. ACL, 2017.
- [Ekm92] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.
- [Gra17] David Graf. Semeval-2017-t8, June 2017.
- [KLA17] Elena Kochkina, Maria Liakata, and Isabelle Augenstein. Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. In *Proc. of SemEval-2017*, pages 475–480. ACL, 2017.
- [LFPR16] Mirko Lai, Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. Friends and enemies of Clinton and Trump: using context for detecting stance in political tweets. In *Proc. of MICAI 2016*, volume 10061 of *LNCS*, pages 155–168. Springer, 2016.
- [MKS<sup>+</sup>16] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proc. of SemEval 2016*, pages 31–41. ACL, 2016.
- [MT13] Saif M Mohammad and Peter D Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.
- [NS13] Nicole Novielli and Carlo Strapparava. The role of affect analysis in dialogue act identification. *IEEE Transactions on Affective Computing*, 4(4):439–451, 2013.
- [OST57] C.E. Osgood, G.J. Suci, and P.H. Tenenbaum. *The Measurement of meaning*. University of Illinois Press, Urbana., 1957.
- [PFB01] James W Pennebaker, Martha E Francis, and Roger J Booth. *Linguistic inquiry and word count (LIWC): LIWC 2001*. Mahway: Lawrence Erlbaum Associates, 2001.
- [PGH<sup>+</sup>13] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. Enhanced senticnet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38, 2013.
- [Plu01] Robert Plutchik. The nature of emotions. *American scientist*, 89(4):344–350, 2001.
- [PVV13] Rob Procter, Farida Vis, and Alex Voss. Reading the riots on twitter: methodological innovation for the analysis of big data. *International journal of social research methodology*, 16(3):197–214, 2013.
- [SSW<sup>+</sup>17] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.
- [VRA18] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
- [Whi09] Cynthia Whissell. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological reports*, 105(2):509–521, 2009.
- [ZLP<sup>+</sup>15] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. Towards detecting rumours in social media. In *AAAI Workshop: AI for Cities*, 2015.
- [ZLP<sup>+</sup>16] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.