Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pages 13–20
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A Multi-Dimensional View of Aggression when voicing Opinion

Arjit Srivastava[1*], Avijit Vajpayee[2*], Syed Sarfaraz Akhtar[3*], Naman Jain[2], Vinay Singh[1], Manish Shrivastava[1]

[1]International Institute of Information Technology, Hyderabad, Telangana, India
[2]Department of Computer Science, Columbia University, New York, NY, USA
[3]Department of Linguistics, University of Washington, Seattle, WA, USA

arjit.srivastava@research.iiit.ac.in
{ssa2184, nj2387}@columbia.edu,
avijitv@uw.edu
vinay.singh@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

The advent of social media has immensely proliferated the amount of opinions and arguments voiced on the internet. These virtual debates often present cases of aggression. While research has been focused largely on analyzing aggression and stance in isolation from each other, this work is the first attempt to gain an extensive and fine-grained understanding of patterns of aggression and figurative language use when voicing opinion. We present a Hindi-English code-mixed dataset of opinion on the politico-social issue of '2016 India banknote demonetisation' and annotate it across multiple dimensions such as aggression, hate speech, emotion arousal and figurative language usage (such as sarcasm/irony, metaphors/similes, puns/word-play).

Keywords: Social Media, Stance, Opinion, Aggression, Hate Speech, Figurative Language, Emotion

1. Introduction

There has been an explosion in terms of the amount of data generated by users online. Social media and online forums encourage users to share their thoughts with the world resulting in a vast resource of opinion-rich data. This has garnered a lot of attention from the research community as it allows for analyzing the interactions between users as well as their usage of informal language in depth.

Stance detection is the task of automatically determining the opinion of users with respect to a given issue. The author of the opinion may be in favour, against or neutral towards the issue. In this paper, we attempt to analyze with respect to stance, nuances of displayed aggression towards supporters / detractors of the opinion as well as the usage of various forms of figurative language such as metaphors, rhetorical questions, sarcasm, irony, puns and word-play. We additionally also look at the emotion arousal level and instances of hate speech. The target issue analyzed in this paper is '2016 Indian banknote demonetisation'. On 8 November 2016, the Government of India announced the demonetisation of all ₹500 and ₹1,000 banknotes of the Mahatma Gandhi Series. It also announced that new banknotes of ₹500 and ₹2,000 banknotes will be circulated in exchange for the demonetised banknotes. However, this decision received mixed reactions from the people of India with many people questioning its effectiveness.

Culpeper (2011) defined verbal aggression as *"any kind of linguistic behaviour which intends to damage the social identity of the target person and lower their status and prestige"* (also cited by Kumar et al. (2018)). Baron and Richardson (2004) identified some characteristics of aggression as :

- Form of behaviour rather than an emotion, motive or attitude.
- Visible intention to hurt or harm (may not be physical).
- Must involve actions / intentions against living beings.
- Recipient is motivated to avoid such treatment.

People often express their opinion on socio-political issues on social media forums like Twitter by displaying aggression towards people that support a contradicting belief or towards particular group of stake-holders on the issue. Given below are some example tweets from our dataset on the target issue of demonetisation in India. The reader is warned on the strongly-worded and derogatory nature of these tweets.

1. **Tweet**: '*ye AAPtards aise behave kar rahe hain jaise Modi ji ne Notebandi nahi inki Nassbandi kara di ho*'

   Translation: '*These AAPtards are behaving as if its not demonetisation but castration for them.*'

   Gloss: "*AAP*": Opposition political party, "AAPtards": slang term for supporters of AAP (inspired by the English slang "Libtards"), "Notebandi": demonetisation of higher currency notes, "Nassbandi": castration

* These authors contributed equally to this work.

Tweet 1 is in favor of the decision and is overtly aggressive towards the people who are voicing their views against it. Due to its abusive language and suggestion of violence, it is also labeled as hate speech. Lastly, this tweet contains word-play as "Notebandi"(Demonetisation) rhymes with "Nassbandi"(Castration).

2. **Tweet**: '*Aam admi se jyada politician ko dikate ho rhi h notebandi se aisa kyun?????*'

   Translation: '*Politicians seem to be more affected by demonetisation compared to the common man. Why is so ?*'

   In tweet 2, the author rhetorically questions why politicians seem to be more troubled about Demonetisation than the normal public, which is to imply that the general public supports the legislation and corrupt politicians are opposing it. This is an example of covert aggression towards the politicians while voicing favourable opinion on the decision.

3. **Tweet**: '*tera kejri mar jaye sala suar to modi ji vaise he ek din ke liye notebandi vapis le lege*'

   Translation: '*If your leader Kejri, a stupid pig, dies then Modi ji would take demonetisation back for a day.*'

   Gloss: "*Kejri*": refering to Arvind Kejriwal (leader of opposition party AAP), "*Modi ji*": Honorific refering to Narendra Modi (Prime Minister of India)

   Tweet 3 supports the decision of demonetisation and is overtly aggressive to both the members of AAP and their leader Arvind Kejriwal. Its author suggests that the opposition leader should die and proceeds to verbally abuse him.

4. **Tweet**: '*what if .. Modi Ji says Mitron ,,, kal raat ko zyada ho gayi thi ,,. Kuch nahi badla he.. #NoteBandi*'

   Translation: '*What if Modi says that he had too much to drink last night and nothing has really changed. #Demonetisation.*'

   Tweet 4 makes a sarcastic joke about how Prime Minister Modi might have been joking and hung over while making this sudden announcement. Despite its humorous take, this tweet is non-aggressive and neutral in stance.

The main contributions of this paper is a unified dataset of 1001 Hindi-English code-mixed tweets annotated for multiple dimensions namely -

- Stance (favourable, against, neutral)
- Aggression (covert, overt, non-aggressive)
- Hate Speech (true, false)
- Figurative language use
  - Sarcasm / Irony / Rhetorical Questions (true, false)
  - Puns / Word-play (true, false)
  - Metaphors / Similes (true, false)
- Emotion arousal (1 to 5 rating)

This is the first attempt at analysing social media opinion on a political issue across varied modalities. More in depth datasets like the one we present here are required for -

- Analyzing the not so apparent forms of verbal aggression displayed on social media.
- Better understanding linguistic patterns when voicing opinion and displaying aggression.
- Analyzing social dynamics of opinion.
- Facilitate classification models that leverage corpora annotated for auxiliary tasks through transfer learning, joint modelling as well as semi-supervised label propagation methods.

The structure of this paper is as follows. In section 2, we review related work in the fields of stance detection, aggression detection, hate speech detection, figurative language constructions, emotion analysis and code-mixed data analysis. In section 3, we explain the annotation guidelines used to for creation of this dataset. Section 4 we present statistics and analysis on the corpus. Finally, section 5 we present our conclusions as well as lay out scope of extending this work.

2. Related Work

User generated data from social media forums like Twitter has attracted a lot of attention from the research community. Mohammad et al. (2017) and Krejzl et al. (2017) analyzed stance in tweets and online discussion forums respectively. The task of stance detection on tweets at SemEval 2016 (Mohammad et al., 2016) led to targeted interest in the area with contributions from Augenstein et al. (2016), Liu et al. (2016) etc. Aggression and offensive language was the focus of a SemEval 2019 task (Zampieri et al., 2019b) and some of the works on aggression identification are Kumar et al. (2018), Zampieri et al. (2019a). Closely related is detecting hate speech in social media which has been explored by Malmasi and Zampieri (2017), Schmidt and Wiegand (2017), Davidson et al. (2017), Badjatiya et al. (2017) among others.

Domains of verbal aggression, abuse, hate have till now been studied in isolation from stance and opinion mining. Additionally, usage of figurative language expressions such as sarcasm, metaphor, rhetorical questions, puns etc.,

** The dataset is publically available at : https://github.com/arjitsrivastava/MultidimensionalViewOpinionMining

when voicing opinion or displaying aggression, has not been explored in depth. As most of the datasets available are annotated with a singular task at hand, it precludes understanding the correlations along multiple dimensions. The next frontier is analyzing in depth these patterns and correlations. To do this, we undertook a data annotation effort on a singular set of tweets for multiple tasks previously studied separately. We hope that this dataset makes way for joint modelling / multi-task learning systems as well provide insights on underlying latent factors.

For this work, we wished to analyze multiple dimensions of opinion on a single target issue. The choice of demonetisation of higher currency notes in India 2017 as our target issue was motivated by the familiarity of the authors and annotators with its nuances as well as the highly polarizing nature of opinions on the topic. Gafaranga (2007) describes code-mixing as use of linguistic units from different languages in a single utterance or sentence and code-switching as the co-occurrence of speech extracts belonging to two different grammatical systems. Majority of user generated data on social media is code-mixed and consequently, so is our dataset. Code mixed datasets for Hindi-English tweets have been previously created for humor (Khandelwal et al., 2018), sarcasm (Swami et al., 2018a), aggression (Kumar et al., 2018), hate speech (Bohra et al., 2018) and emotion (Vijay et al., 2018).

3. Annotation

Swami et al. (2018b) had collected 3500 code-mixed Hindi-English tweets using the Twitter Scraper API filtering by the keywords "*notebandi*" and "*demonetisation*" over a period of 6 months after Demonetisation was implemented and annotated them for stance (favourable, against and neutral). We randomly sampled 1001 tweets from this dataset and annotated these sampled tweets for the dimensions (3 domain expert annotators for each dimension) :

- Aggression : Overt vs. Covert vs. Neutral
- Hate Speech : True vs. False
- Sarcasm / Irony / Rhetorical Question : True vs. False
- Metaphor / Simile : True vs. False
- Pun / Word-play : True vs. False
- Emotion Arousal : 5 point ordinal scale

The final label on each binary classification dimension was taken as the majority label from choices of 3 annotators. For aggression classification, which was a multi-class classification, adjudication was provided by us for cases where no simple majority could be reached. For emotion arousal levels, scores from individual annotators were averaged for the final emotion arousal level score.

We also re-annotated the original dataset for stance for it had favourable or against tags only on tweets that displayed outright support or disapproval respectively. We found that majority of opinion was displayed through attacking / supporting other opinions on the issue i.e. examples of indirect or implied support / disapproval. For example look at the tweet below -

5. **Tweet**: '*Notebandi k khilaf kyu ho...? Kaale dhan m share holder ho kya @ArvindKejriwal*'

   Translation: '*Why are you against demonetisation ? Are you are shareholder in black money @ArvindKejriwal*'

   Gloss: "*kale dhan*": black money, "*Arvind Kejriwal*": Leader of opposition political party AAP, "*Notebandi*": demonetisation

   Tweet 5 was originally classified as a neutral stance. We feel that cases like above can be confidently annotated as favourable to the issue (i.e. favourable to demonetisation). The author rhetorically and sarcastically questions the opinion, intentions and reasons of those against the issue (in this case leader of opposition party). This tweet is also an example of what we consider covert aggression.

For aggression annotation, we follow the guidelines by Kumar et al. (2018) who had presented a detailed typology of aggression on Hindi-English code-mixed data. We only annotate for aggression level and they had additional layers based on discursive role (attack, defend, abet) and discursive effect (physical threat, sexual aggression, gendered aggression, racial aggression, communal aggression, casteist aggression, political aggression, geographical aggression, general non-threatening aggression, curse). The definitions for 3 aggression levels along with examples from our dataset are :

Covertly-Aggressive (C) Contains text which is an indirect attack and is often packaged as (insincere) polite expressions (through the use of conventionalized polite structures) such as satire, rhetorical questions, etc.

6. **Tweet**: '*Notebandi ka niyam : khata nahi hai to khulwao. Aam aadmi : khulwa to lun. Par bhai bank main ghusun Kasey ?*'

   Translation: '*Rule of Demonetisation: If you don't have an account then open one. Common man: I'll open but let me know how to enter the bank first?*'

   Disapproval of demonetisation through sarcastic reference to long queues in front of banks due to high demand for exchange of demonetised currency.

Overtly-Aggressive (O) Contains texts in which aggression is overtly expressed either through the use of specific kind of lexical items, syntactic structures or lexical features.

7. **Tweet**: '*Ye Notebandi Atankbaadiyo aur Bharashtachaariyo ki NAKEBANDI hai. Sare Rashtrabhakta is nakebandi ke sath aur samarthan me aye.*'

   Translation: '*Demonetisation is a barricading of terrorists and corrupt. All the nationalists should support this barricading.*'

Non-Aggressive (NAG) Refers to texts which are not lying in the above two categories.

8. **Tweet**: '*kya Aam aadmi ke liye NoteBandi ka Faisla Shi hai?*'

   Translation: '*Is the decision of demonetisation in the favour of common man?*'

Prior works regarding sarcasm and irony detection on social media data like Reddit (Wallace et al., 2014) and Twitter (Bamman and Smith, 2015) have shown that context is essential in understanding sarcasm. Therefore, most social media datasets of sarcasm are self-annotated i.e. hashtag specific twitter scraping like #sarcasm and #notsarcasm. As we are re-annotating a previously scraped dataset which was not self-annotated through specific hashtags, we rely on the domain knowledge of context expert annotators on the Indian socio-political scenario and focus issue of demonetisation. This however is not a drawback because in a dataset like ours , rich with strongly opinionated tweets, annotating sarcasm is fairly easy. In the current scope of the research, rhetorical questions are thought of as functioning similar to sarcasm and irony. We understand that fine grained linguistic differences between sarcasm, irony and rhetorical questions exist, for our purpose we have clubbed them into a single category of figurative language. Similarly, puns and word-play are merged into a single category of figurative language as well and the annotation guidelines were based on the SemEval 2017 task of detecting english puns (Miller et al., 2017). Rhyming usage of 'Notebandi' (demonetisation) with 'Nasbandi' (castration) as shown in the earlier examples, was the most common word-play seen. A third figurative language category of metaphors (and occasionally similes) can also be clearly observed in our corpus. Metaphor identification has been typically treated as a token level or phrase level tagging task (Shutova and Teufel, 2010). To be consistent we other figurative language categories used in this work, we annotated metaphors at the tweet level which was also the annotation level for SemEval 2015 task on figurative language in Twitter data (Ghosh et al., 2015). The following tweet is an example of metaphor usage -

9. **Tweet**: '*kabhi kabhi sher ka shikar karne ke liye bhed (aam janta) ko chara banana padta hai. notebandi*'

   Translation: '*Sometimes sheep need to be sacrificed in order to to hunt lions Demonetisation.*'

   In tweet 9, 'sheep' is a metaphor for some members of common public and 'lions' is a metaphor for large scale corruption.

Burnap and Williams (2015) defined hate speech as responses that include written expressions of hateful and antagonistic sentiment toward a particular race, ethnicity, or religion. They used a binary classification scheme of hate speech vs. non hate speech, which was also followed by Bohra et al. (2018) for their dataset on Hindi-English code-mixed tweets. Malmasi and Zampieri (2017) used a 3 way classification scheme between hate speech vs. offensive language but not hate speech vs. no offensive language. As aggression levels are highly predictive of offensive language but not of hate speech category, we used a binary classification speech. However annotators faced difficulty in differentiating over a personal attack full of hatred than a community being targeted. An example :

10. **Tweet**: '*ab itni taklif hai to atmadaah kyo nahi kar lete notebandi k khilf. Delhi walo ko bhi mukti milegi tumse*'

    Translation: '*If you have such a huge issue with it, why don't you perform a self-immolation? The people of Delhi would also get freedom from you*'

    In tweet 10, the author is referring to Arvind Kejriwal who is the leader of opposition party AAP and also the Chief Minister of Delhi (capital of India). The author suggests Kejriwal should kill himself to free the residents of Delhi. In the process of supporting the decision of Demonetisation, the author of the tweet is making extreme and graphic suggestions towards one of the main opponents of target issue.

Emotion classification in text is widely understood as lying across two orthogonal dimensions - valence (polarity of emotion) and arousal (intensity of emotion) (Russell and Barrett, 1999). Despite that, many works on emotion classification in text have generally used directly annotated 6 emotion categories (happy, sad, anger, fear, disgust, surprise) instead of first annotating arousal and valence separately before mapping them into emotion categories. We restricted the scope for this project to analyze only for emotion arousal level as emotion valence level is analogous to sentiment. For emotion arousal level, Bradley and Lang (1999) averaged annotations on a 9 point scale and Mohammad (2018) used a Best-Worst scale to obtain fine-grained scores. Similar to the SemEval 2017 task (Rosenthal et al., 2019) for sentiment analysis on Twitter, we use a 5-point ordinal scale (Very Low, Low, Neutral, High, Very High) for emotion arousal level.

4. Data Statistics and Analysis

Table 1 presents the tweet level average statistics on the corpus. The dataset tweets contain majorly Hindi language

tokens (written in Roman script instead of Devnagri). A total of 119 tweets had discernible code-mixing (3 or more english words). As our tweets were sampled from the dataset by Swami et al. (2018b) who had referred to their dataset as code-mixed, we continue to refer it that way. Subsequent model building on this corpus would benefit from special handling for token-level spelling differences that come with Devnagri to Latin script switching for Hindi.

Table 2 has the corpus wide statistics across various phenomena annotated. There is a significant skew towards favourable stance in the corpus. To accommodate for this imbalance, subsequent analysis of phenomena with respect to stance contain marginal class percentage statistics, for example percentage of sarcastic tweets in favour of the issue with respect to total number of tweets favourable to the issue. Another point to note is the very low number of hate speech instances. This could be attributed to the stringent guideline that only directed abusive attacks on specific groups/communities are to be regarded as hate speech. Annotations with looser guidelines, where personal offensive language against individuals are also considered hate speech, would correlate highly with overt aggression category. Since we annotated on tweets regarding a polarizing legislation, it was expected that a fair amount would display aggression (either covert or overt). The same observation is evident from the statistics.

Avg. # tokens	21.1
Avg. # tokens (EN)	1.0
Avg. # tokens (HI)	16.9
Avg. # tokens (Rest)	3.2

Table 1: Tweet Level Statistics

4.1. Annotation Agreement

We used Fleiss's kappa to measure inter-annotator agreement on categorical annotation tasks and the results are given in table 3. Due to the clear polarizing nature of issue at hand, annotations for stance were of very high correlation. Hate speech annotations had the worst kappa score and can be attributed to what constitutes a personal abusive attack. For figurative language use, the annotations for puns and word-play were of higher correlation as can be expected due to the apparentness in surface forms. Annotations for sarcasm / irony / rhetorical questions while still being of high agreement had lower agreement rate than both metaphors / similes as well as word-play. This can be attributed to the general greater subjective nature of sarcasm as well as it being a more context-dependent phenomenon than metaphor or word-play.

Task	Category	# Tweets
Stance	Favour	583
	Against	180
	Neutral	238
Aggression	Overt	140
	Covert	264
	None	597
Hate Speech	True	29
Figurative Language	Sarcasm / Irony / Rhetorical Ques.	163
	Word-play / Pun	140
	Metaphor / Simile	189

Table 2: Distribution of annotations across corpus

Task	Fleiss's kappa
Stance	0.84
Aggression	0.62
Hate Speech	0.47
Sarcasm / Irony / Rhetorical Questions	0.61
Puns / Word-play	0.72
Metaphors / Similes	0.65

Table 3: Fleiss's kappa score on multiple annotations across dimensions

Spearman's Rank Correlation Emotion Arousal		
Annotator	2	3
1	0.655	0.652
2		0.64

Table 4: Spearman correlation on emotion arousal annotations across annotator pairs

Table 4 gives the Spearman's rank correlation coefficient across 3 annotators for emotional arousal which has been rated on an ordinal scale of 1 to 5. Although annotating for emotion is a fairly difficult task and annotating for only the arousal dimension even more so. However, we achieve a decent average correlation of 0.65 which can be attributed to the fact that these tweets were sampled for a polarizing issue which had clearly apparent emotional states (high arousal emotions like anger as well as low arousal emotions like sadness). For each pair of annotators, the results of emotional arousal agreement were statistically significant with p-values $<<<$ 0.005.

4.2. Stance specific analysis

Table 5 presents the statistics of hate speech across stance classes. An anomalous observation is the higher marginal percentage of hate speech evidence for *neutral* stance. This could be attributed to the poorer understanding of what constitutes hate speech. Additionally, upon investigating we found tweets similar to the one given below. Though the tweet does not take a definitive stance on the issue at hand (demonetisation), it is an abusive personal attack at an individual as well as a group.

11. **Tweet**: '*MR. RAVISH VYAPARI IMAANDAR HAI.KANOON KA SANMAAN KARTSHAI. PAR*

MEDIA NEWS AUR TV SAB SAALE CHOR AUE HARAMKHOR HAI. NOTEBANDI'

Translation: '*Mr. Ravish, businessmen are honest and respect the law. But media, news and TV (personalities) are thieves and bastards.*'

Glosses: "*Ravish*": Refering to news anchor Ravish Kumar

Tweet 11, defends integrity of businessmen while attacking and name calling news personalities.

Stance	Marginal Class % of Hate Speech
Favour	2.92%
Against	2.2%
Neutral	3.36%

Table 5: Distribution of hate speech across stance

Table 6 gives the distribution of aggression categories (*covert / overt / non*) across stance. It is interesting to note the comparisons for overt vs. covert aggression when in favour (majority population stance in this sample) as opposed to against (minority population in this sample) on the issue. Although covert aggression evidence is always more than overt aggression evidence across stance categories, the difference is much lesser for *favourable* stance samples. It is not difficult to hypothesise that holding a majority stance on issues will lead to open bullying in a lot of cases. Users in minority tend to be more covert to possibly avoid being bullied by the majority group. Though validating this social hypothesis based on analysis of multiple issues is beyond our current scope.

Table 7 presents the distributions of figurative language use across stance classes. It is evident from the data of *against* issue category, the usage of all types of figurative language is consistently high. It should also be noted that evidence for sarcasm is especially higher in *against* issue opinion (minority stance in this dataset). Keeping in mind the observations on covert aggression when voicing minority stance, it can be noted that covert aggression is expressed through figurative language like sarcasm and puns. Metaphors are not as disguised as sarcasm and puns and we see that it does not follow the same pattern with respect to stance. The scope of this work is limited to a single issue and it would be interesting to note if these trends are observed across datasets. A dataset of annotations of multiple issues would allow for hypothesis testing to validate these trends.

Stance	Aggression	Marginal Class % Aggression
Against	Overt	8.3%
	Covert	40%
	None	**51.7%**
Favour	Overt	17.8%
	Covert	23.8%
	None	**58.3%**
Neutral	Overt	8.8%
	Covert	22.3%
	None	**68.9%**

Table 6: Distribution of aggression across stance

Finally in table 8, statistics for emotion arousal are presented across stance classes. Opposed to prior analyzed phenomena (hate speech, aggression and figurative language use), the data for emotion arousal is ordinal on a 1 to 5 scale. The average emotion arousal for *favourable* stance (majority class) is much more than that in *against* stance (minority class). Similarly, looking at the very high arousal state bucket of *5* emotion arousal (when all three annotators gave a 5 rating), the percentage for majority stance (favourable) is three times than that for minority stance (against). These findings are in line with the observations for other phenomena like overt aggression and figurative language use in the majority stance. The higher percentage of lowest arousal state tweets when against the issue must also be noted. These lowest arousal tweets correspond to emotions like depression and sadness.

5. Conclusion and Future Work

This research was motivated by the need to provide a ground-work for analysis of the nuances of opinion on social media with respect to aggression and figurative language use. The observed correlations are encouraging and call for a deeper analysis of these social dynamics. Testing for statistical significance along with corpus-linguistic analysis of informative words for each category was beyond our current scope. The first aim would be to create similar corpora on wide variety of issues (not limited to political debate) to evaluate the consistency of these trends and determine significance of our findings.

Though the scope of this project was limited to corpus creation and analysis of interactions across phenomena, the larger goal is to allow for better classification systems on social media data. An immediate goal is to build baseline models and analyze their performance on the different phenomena annotated in this corpus. It would be interesting to compare performance of models that directly model a single dimension with those models that have cascaded or joint modeling on multiple dimensions. Another avenue we would like to explore is semi-supervised label propagation utilizing both larger corpora on a single dimension such as sarcasm as well as this corpus containing multi-dimensional annotations. Having a single corpus of annotations across dimensions has allowed the possibility to explore transfer learning strategies in classification.

For the sake of keeping this breadth-wise annotation effort manageable, we annotated for a 1001 tweets. We plan to extend this dataset to all 3500 tweets from the original dataset created by Swami et al. (2018b). We further plan to annotate these tweets for named entities as well as 6 emotion classes similar to Vijay et al. (2018).

Stance	Sarcasm / Irony / Rhetorical Question		Pun / Word-play		Metaphor / Simile	
	Raw Count	Marginal Class %	Raw Count	Marginal Class %	Raw Count	Marginal Class %
Against	45	**25%**	30	**16.7%**	35	19.4%
Favour	76	13%	84	14.4%	123	**21.1%**
Neutral	42	17.6%	26	10.9%	31	13%

Table 7: Distribution of figurative language across stance

Stance	Marginal Class %					Emotional Arousal Class Avg.
	1-2	2-3	3-4	4-5	5	
Favour	3.6%	23.67%	49.91%	18.01%	4.8%	3.26
Against	13.3%	26.67%	47.78%	10.56%	1.67%	2.91
Neutral	10.9%	36.97%	44.12%	6.72%	1.3%	2.83

Table 8: Marginal distribution of emotional arousal across stance

6. Bibliographical References

Augenstein, I., Rocktäschel, T., Vlachos, A., and Bontcheva, K. (2016). Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

Bamman, D. and Smith, N. A. (2015). Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Baron, R. A. and Richardson, D. R. (2004). *Human aggression*. Springer Science & Business Media.

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., and Shrivastava, M. (2018). A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of peopleâs opinions, personality, and emotions in social media*, pages 36–41.

Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology â.

Burnap, P. and Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Culpeper, J. (2011). Impoliteness: Using language to cause offence. *Impoliteness: Using Language to Cause Offence*, pages 1–292, 01.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Gafaranga, J. (2007). 11. code-switching as a conversational strategy. *Handbook of multilingualism and multilingual communication*, 5(279):17.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., and Reyes, A. (2015). Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478.

Khandelwal, A., Swami, S., Akhtar, S. S., and Shrivastava, M. (2018). Humor detection in english-hindi code-mixed social media content: Corpus and baseline system. *arXiv preprint arXiv:1806.05513*.

Krejzl, P., Hourová, B., and Steinberger, J. (2017). Stance detection in online discussions. *CoRR*, abs/1701.00504.

Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018). Aggression-annotated corpus of hindi-english code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Liu, C., Li, W., Demarest, B., Chen, Y., Couture, S., Dakota, D., Haduong, N., Kaufman, N., Lamont, A., Pancholi, M., et al. (2016). Iucl at semeval-2016 task 6: An ensemble model for stance detection in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 394–400.

Malmasi, S. and Zampieri, M. (2017). Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

Miller, T., Hempelmann, C., and Gurevych, I. (2017). SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada, August. Association for Computational Linguistics.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3), June.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Rosenthal, S., Farra, N., and Nakov, P. (2019). Semeval-

2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.

Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Shutova, E. and Teufel, S. (2010). Metaphor corpus annotated for source-target domain mappings. In *LREC*, volume 2, pages 2–2.

Swami, S., Khandelwal, A., Singh, V., Akhtar, S. S., and Shrivastava, M. (2018a). A corpus of english-hindi code-mixed tweets for sarcasm detection. *CoRR*, abs/1805.11869.

Swami, S., Khandelwal, A., Singh, V., Akhtar, S. S., and Shrivastava, M. (2018b). An english-hindi code-mixed corpus: Stance annotation and baseline system. *CoRR*, abs/1805.11868.

Vijay, D., Bohra, A., Singh, V., Akhtar, S. S., and Shrivastava, M. (2018). Corpus creation and emotion prediction for hindi-english code-mixed social media text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 128–135.

Wallace, B. C., Kertz, L., Charniak, E., et al. (2014). Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.