Exploring the Diversity of Opinions on Affirmative Action Through Extended Stance Detection Among Reddit Users

Anonymous ACL submission

Abstract

Affirmative Action (AA) has remained a controversial topic in the U.S. for several decades. While previous research has extensively explored AA from legal, social, and ethical aspects, there is a lack of computational work to investigate this topic from the lens of users on social media. By collecting over 2,500 posts from 23 prominent Reddit communities, our study attempts to gain a better understanding of how online users view AA. We build upon the previous work on stance detection, by introducing a new set of stance categories that can efficiently reflect a diverse range of opinions on AA. Finally, we explore the performance of three LLM-based classifiers, powered with four carefully designed prompts to classify these categories and run several analyses to enhance our understanding of the AA discourse. Our findings show that GPT-4 with instruction and examples is the most efficient for classifying stance. We found opposition towards AA more common than support in our dataset with discussions on college admission, and race more prevalent. This study enhances the field by presenting a novel dataset of stances derived from Reddit data and initiates a conversation on broadening binary evaluations of viewpoints on controversial subjects.

1 Introduction

Affirmative Action (AA) represents a set of policies and practices aimed at addressing historical and ongoing inequalities in employment, education, and other sectors by providing opportunities to historically marginalized groups. These measures are designed to promote diversity and rectify socio-economic disparities caused by past discrimination, ensuring that everyone has a fair chance to succeed regardless of their background (Thomas, 1990). Since the implementation of AA in 1961, however, these policies have been faced with a variety of public opinions from various stakeholders (Cahn, 2013; Kennedy, 1985). For instance, while (Crenshaw et al., 1995; Rowell and Williams, 1996) contend that AA is a right step in the right direction to mitigate racial inequities and help remove barriers for marginalized groups, studies like (Fish, 2000; Chemerinsky, 1996) question its effectiveness or conflicts with a merit-based system.

A large body of qualitative work has been conducted in an attempt to gauge public opinion on affirmative action across different demographic groups (Pew Research Center, 2023). For example, findings of a survey work (Bowman and O'Neil, 2016) indicate that while Americans generally support increased diversity in colleges, they are less favorable towards race or ethnicity-based admissions. Additionally, the survey shows varying levels of support for AA among different racial groups. In this work, we take a computational approach to complement these qualitative studies. More specifically, we use stance detection to study the wide range of positions from Reddit users on AA.

Stance detection has been extensively used to study topics that drive controversy and conflicting views such as abortion, feminism, climate change, and politics (Mohammad et al., 2016). Using stance detection techniques enhances understanding of diverse viewpoints and public opinion dynamics on controversial issues. By analyzing the stances taken by individuals or groups, we can uncover belief patterns and sentiments and track changes in public opinion, which offers insights into how and why perspectives shift over time. One shortcoming with the widely used stance labels, i.e., Favor or Against, is that they do not always sufficiently capture the variety of positions taken by the speaker toward the target, falling short to achieve a full understanding of the nuances in stance analysis (Liang et al., 2024; Xu et al., 2022; Hardalov et al., 2021; Simaki et al., 2020). To address this shortcoming, we propose a new set of stances to represent the diversity of positions in discussions related to AA. Based on the literature and after multiple iterations we break down Favor and Against into their respective strong or weak versions, we introduce Skeptical as a new label and add Question, Question Favor, and Questions Against, as new stances to analyze public opinion on AA. We demonstrate the applicability of these new stances in providing a more comprehensive understanding of public views toward affirmative action.

Our data is collected from a diverse range of 23 subreddits over a span of 13 years (from 2010 to 2022). Using the new stance labels, we manually annotated 400 posts from Reddit and used different LLMs with 4 prompt strategies to find the most efficient model for labeling 2,548 posts in our dataset. Our results confirm the usefulness and application of this new set of stance categories in broadening our understanding of AA discourse. Our *Fewshot + Instruction + Definition* LLM-based classifier using GPT-4 achieved 0.66 for F1 score, which is competitive to the state-of-the-art, given the complexity of the stance detection task and the significant addition of categories in this study.

This work contributes to stance analysis research by 1) Introducing a new set of stance labels that enhance our ability to capture the online discourse and will improve the efficiency of stance detection task 2) Releasing a new affirmative action dataset, collected from Reddit, and 3) using LLM-based classifiers using different prompts to achieve competitive stance classifiers with new stance labels.

2 Related Work

2.1 Affirmative Action

Affirmative Action, also referred to as Positive Discrimination, policies started in the 1960s as a response to the needs and demands of those disproportionally impacted by racial bias and discrimination (Crenshaw et al., 1995). AA takes steps to promote equality in opportunities and outcomes and provide legal approaches that are designed to level the playing field for historically marginalized communities and underrepresented populations that have been deprived of equal access to opportunities (Rowell and Williams, 1996). While this approach is often considered "color blind," since the implementation of AA policies, these programs have been the subject of a nationwide debate over their merit, necessity, and legal base (Chemerinsky, 1996; D'Souza and Edley Jr, 1996; Coleman, 1999); some American sociologists argue that it falls short in achieving true racial equality in outcomes. Bengtson (2024) highlights mitigating discrimination, equality of opportunity, and diversity promotion as the three prominent arguments in favor of AA based on a relational egalitarian theory of justice. On the other hand, Fish (2000) contends that AA compromises merit-based admissions, which can lead to the admission of students who may not meet the same academic standards as their peers. This could result in a mismatch where students struggle to keep up academically, ultimately harming their educational experience and outcomes. In June 2023, the U.S. Supreme Court ruled on the significant case of "Students for Fair Admissions v. Harvard," declaring that race-conscious admissions policies at colleges violate the Equal Protection Clause of the 14th Amendment (Supreme Court of the United States, 2023).

2.2 Stance Analysis

Stance detection is defined as the task of automatically identifying the ideological position (e.g., favor, against, or neutral) toward a specific target, usually a controversial topic (Kucuk and Can, 2020; Hardalov et al., 2022; AlDayel and Magdy, 2021; Lai et al., 2020). With the emergence of social media platforms as the primary source of information and channel of communication, these online environments have become a convenient place for people to share and discuss their viewpoints toward various topics such as legalization of abortion, climate change, or election (Haddington, 2006). With the increasing access to publicly available online datasets, a growing number of studies have attempted to analyze and understand people's attitudes, opinions, and behavior (Boyd et al., 2015; Park et al., 2015). Furthermore, recent developments in various areas such as NLP facilitated the task of modeling and analyzing people's polarized views and stances (Mohammad et al., 2016, 2017), emotion (Wiebe et al., 2005), personality (Park et al., 2015), and moral values (Graham et al., 2013; Sagi and Dehghani, 2014).

Sobhani et al. (2015) used topic modeling for stance detection and classification in online news comments. Mohammad et al. (2016) introduced the SemEval 2016 task, which became a baseline for many stance studies in user-generated texts. Using Twitter (a.k.a X now) as the data source, they created labeled datasets on a range of controversial topics like feminism, abortion, atheism, and climate change. Researchers have combined a range of linguistic features such as sentiment (Mohammad et al., 2017), and moral analysis (Rezapour et al., 2019, 2021), as well as network features such as user following, follower, and retweet information (Aldayel and Magdy, 2019) to improve the efficiency and explainability of stance analysis. Such approaches have led to several multi-task studies (e.g., stance + sentiment analysis), collectively enhancing our understanding of public opinions on social media and beyond (Li and Caragea, 2019; Sobhani et al., 2016). To improve stance detection, Popat et al. (2019)) used the transformer-based BERT model. The result of their experiments on a benchmark dataset called Perspectrum (Chen et al., 2019) demonstrated the success of their approach in increasing the performance of stance classification compared to state-of-the-art models.

2.3 LLMs and Stance

The recent rise of LLMs has resulted in a new body of research, aiming to incorporate these models in efficiently studying and analyzing stance and related areas such as moral values (Shamik Roy, 2022; Jiang et al., 2022; Roy and Goldwasser, 2021; Sharma et al., 2023; Kang et al., 2023) Shamik Roy (2022) used in-context learning, with zero-shot or few-shot learning, to generate outputs that can predict the moral dimension of tweets or the moral role of different entities in the tweet. Roy and Goldwasser (2021) investigated how prompts can be beneficial in comparing different ways of framing issues between various media outlets while Jiang et al. (2022) examined the ability of an LLM-based framework, with fine-tuned prompts on each community's dataset, in predicting the favorability of political figures between two ideological communities. Sharma et al. (2023) fine-tuned different text-based LLMs such as XLNet (Yang et al., 2019), XLM-RoBERTa, and Transformer XL for stance prediction on tweets about gun control and abortion, demonstrating high precision and recall. An ensemble of these models achieved the highest performance, indicating that combining outputs from multiple LLMs improves accuracy. Kang et al. (2023) introduced the Value Injection Method (VIM), an approach that fine-tunes LLMs to align with specific human values, enhancing their ability to predict opinions and behaviors. Using value distributions, models can generate arguments and answer questions that reflect these values, outperforming baseline models in stance prediction tasks.

2.4 Stance beyond binary

While most researchers have approached stance detection as a binary problem (pro/con or favor/against), stance tends to be more complex, context-dependent, and impacted by the intensity of one's opinion (Xu et al., 2022; Hardalov et al., 2021; Liang et al., 2024; Simaki et al., 2020). Hardalov et al. used various datasets to illustrate that binary classification often fails to capture the full spectrum of stances, particularly in diverse social media contexts. For example, stances on rumors might include categories like endorse, deny, or question, which are not well-represented by support, oppose, or neutral labels.

Simaki et al. (2020) adopted a cognitive-functional framework that identifies ten stance categories. They used agreement/disagreement, certainty, contrariety, hypotheticality, necessity, prediction, source of knowledge, tact/rudeness, uncertainty, and volition as stance categories and proved their usefulness in the annotation of a Brexit corpus. Qazvinian et al. (2011) proposed Believe/Endorsement and Deny/Doubtful/Neutral as two useful labels for the rumor classification and misinformation detection task. Our study builds upon these works by expanding the number of stance categories used in annotating and classifying affirmative action posts.

3 Method

3.1 Data Collection

To collect data for our study, we first used the two most relevant keywords to AA, i.e., "affirmative action" and its widely used equivalent in other countries "positive discrimination" as our search queries on Reddit, and found 63 subreddits that had at least one post related to AA. Finally, after manually reviewing the 63 Reddit communities, we filtered out those that were not directly relevant to the scope and context of our study and chose 23 subreddits that primarily discuss politics and political ideologies, such as *r/AskALiberal* and *r/askaconservative*, and social issues, such as *r/sociology*, and popular subreddits known for general conversations and the exchange of ideas, such as *r/AskReddit*. Table 4 shows the complete list of selected subreddits in our data.

We then used Pushshift (Baumgartner et al., 2020) to collect data from these subreddits spanning from January 1, 2005, to December 30, 2022. We filtered the posts discussing affirmative action using our chosen keywords ("affirmative action" and "positive discrimination") from all the identified subreddits. After excluding posts without bodies, author-deleted posts, and posts that were too short or too long (based on the 10th and 90th percentile of post length), we ended up with 2,548 posts in our dataset.

3.2 Extension of Stance Beyond Binary

Stance analysis in academic and social discourse has traditionally employed three categorical labels: Favor, Against, and Neutral. However, these classifications often fall short of capturing the full complexity of human perspectives, which can exhibit partial agreement or disagreement, conditional responses, or skepticism. Such limitations highlight the need for a more nuanced understanding of discourse, particularly in areas as controversial as AA. Researchers categorize stance into several types (Jaffe, 2009; Couper-Kuhlen and Selting, 2017; Andries et al., 2023). The affective stance deals with emotions or attitudes towards a subject. For example, a person might express a strong emotional support for AA by stating, "I feel very strongly that affirmative action is essential for achieving true equality." The epistemic stance reflects beliefs or levels of certainty, such as in the assertion, "I think affirmative action effectively promotes diversity in educational institutions." Lastly, the deontic stance concerns obligations or permissions, e.g., "Universities must implement affirmative action policies to ensure fair admission practices."

In a complex discourse, especially on social media, public debates, and scholarly discussions, the richness of human communication is often lost with simplistic categorizations. By adopting a broader range of labels, researchers can capture more accurately the degrees of agreement or disagreement, skepticism, ambivalence, and nuanced questioning present in discourse. This enhanced categorization facilitates a deeper, more precise analysis of textual data, crucial across various fields.

We followed a systematic approach to expand stance labels in this study in relation to discussions on AA. First, we conducted a literature review with the purpose of identifying different categories of stance labels used to analyze public discourses (Liang et al., 2024; Xu et al., 2022; Hardalov et al., 2021; Simaki et al., 2020). Taking a data-driven approach, we reviewed a random sample of posts to take a deeper look into the linguistic variations, tone and diverse range of ideas and arguments used in the context of AA. This investigation led to the development of nine stance categories better suited to our dataset, including 'Strong Favor', 'Strong Against', 'Weak Favor', 'Weak Against', 'Question', 'Question Favor', 'Question Against', 'Skeptical', and 'No Stance'. These categories were refined through iterative coding and discussions among researchers, followed by testing annotations to confirm their relevance and applicability.

3.3 Data Annotation

Three annotators (with diverse backgrounds: an international student, an African-American student, and one Asian-American student with data/computer science and psychology backgrounds and high familiarity with AA) annotated a randomly sampled data of 400 posts using the nine newly developed stance labels. We used Zooniverse's interactive environment [1] for the annotating task. The sample posts were split into 3 subsets of 100, 200, and 100. After each round, disagreements were discussed and resolved to reach an acceptable level of agreement between annotators. We used Fleiss Kappa to measure the degree of inter-rater reliability between the three annotators. We achieved an average of 0.46 Kappa agreement rate, an acceptable range given the complexity of stance detection task and the significant increase in the number and diversity of stance categories used in this work.

3.4 Zero- and Few-shot Classification of Stance

We used three LLMs, namely OpenAI's GPT-3.5 and GPT-4 [2]and Mixtral-8x7B (Jiang et al., 2024) for the classification of stance in our dataset. We designed four prompting strategies to guide the LLMs in the classification tasks:

- *Zeroshot + Instruction (Z+I)*, where only a general instruction for the stance classification task has been provided to the model.

- *Zeroshot + Instruction + Definition (Z+I+D)*, where the definitions of stance analysis and different stance categories are added to the instruction.

- *Fewshot + Instruction (F+I)*, where multiple example-label pair per stance category is added to the general instruction.

[1]https://www.zooniverse.org/
[2]https://openai.com/

- *Fewshot + Instruction + Definition (F+I+D)*, where the definitions of stance analysis and different stance categories, as well as example-label pairs per categories, are included.

Appendix D shows the prompts used for each of these strategies. We split 400 annotated posts into two subsets of 50 validation posts for experimenting with the prompts and 350 posts for evaluating the performance of these models. Different versions of GPT models such as GPT-4-turbo, GPT-4-0125-preview, GPT-4-turbo-preview, GPT-3.5-turbo-0125, GPT-3.5-turbo were tried on the validation set to choose the best version of each model for classification. We set the temperature to zero in all LLMs to ensure consistency.

3.5 Keyphrase and Theme Analysis

To enhance our understanding of the nuances in Reddit's discourse about AA, we leverage keyphrase extraction to analyze what terms or phrases have been most prevalent in the posts holding different types of stances. To do so, we use the KeyBert package (Grootendorst, 2020), a keyword extraction tool that uses BERT embeddings (Devlin et al., 2018) to create keywords and keyphrases most similar to a text [3]. Our data and codes will be shared upon the paper's acceptance.

4 Results

Data Analysis. Table 1 shows the number of posts annotated for each stance category. While the majority of annotated posts are either holding a neutral position (No Stance) or asking questions about affirmative action, a significant number of posts take a stance against AA (113 for strong, weak, and question against posts combined) compared to a smaller share of posts that are supporting AA (25 for all categories combined).

Classification. Table 3 shows the performance of LLM-based classifiers used with four types of prompts strategies for stance classification on the test set of 350 posts. As shown in the table, GPT-4 has consistently achieved the highest performance (in terms of precision, recall, and F1) across different prompt types compared to other models. Furthermore, it is evident that the performance of GPT-4 and Mixtral have incrementally increased as we have gradually added to the complexity of our prompts by providing further definitions and examples, but for GPT-3.5 the best performance is achieved when zeroshot prompt with only Instruction and Definition is used. The best performance overall is achieved with GPT-4 and (F+I+D) pair, with F1 score of 0.66.

[3] https://maartengr.github.io/KeyBERT/

Stance Category	#Posts (Annotated)	#Posts (All)
No Stance	159	666
Question	76	382
Strong Against	45	465
Weak Against	35	211
Question Against	33	232
Skeptical	27	311
Strong Favor	11	96
Question Favor	8	39
Weak Favor	6	146

Table 1: Number of each stance label in our annotated as well as all posts

Subreddit	#.Post	Top Stance
unpopularopinion	416	SA, NS, WA
AskReddit	399	Q, QA, NS
changemyview	394	SA, S, NS
ApplyingToCollege	341	NS, Q, WF
Libertarian	168	NS, SA, WA
AsianMasculinity	143	NS, SA, WA
AskALiberal	143	NS, Q, QA
NoStupidQuestions	92	Q, QA, NS
TrueUnpopularOpinion	83	SA, NS, SF
Destiny	68	NS, S, Q
asianamerican	59	NS, Q, SA
neoliberal	59	NS, Q, S
TooAfraidToAsk	46	Q, QA, S
AskSocialScience	43	Q, NS, QA
askaconservative	40	Q, NS, QA

Table 2: List of top 15 subreddits with the top three frequent stances for each subreddit. SA: Strong Against, WA: Weak Against, SF: Strong Favor, WF: Weak Favor, Q: Question, QA: Question Against, QA: Question Favor, S: Skeptical, NS; No Stance

Error Analysis. To further analyze the misclassification errors, we compare the ground truth labels of the test set against the labels generated by the best LLM-Prompt pair (GPT-4 + (F+I+D)). Our analysis shows that most of the errors have occurred for No Stance label, where it has been misclassified with Skeptical, Weak Against and Weak Favor, suggesting that distinguishing a neutral position from a weak stance or skepticism can be challenging for LLM classifiers.

For instance, *'Instead of AA, we should make it illegal to ask about name/sex/race on job applications.'* was tagged as Skeptical instead of No Stance and *'All men are NOT created equal When*

Prompt Type	Model	Precision	Recall	F1
	GPT-3.5	0.51	0.27	0.28
Zeroshot + Instruction	GPT-4	0.58	0.39	0.40
	Mixtral	0.43	0.19	0.14
	GPT-3.5	0.58	0.37	0.40
Zeroshot + Instruction + Definition	GPT-4	0.67	0.49	0.51
	Mixtral	0.54	0.23	0.19
	GPT-3.5	0.56	0.28	0.3
Fewshot + Instruction	GPT-4	0.65	0.57	0.59
	Mixtral	0.57	0.32	0.28
	GPT-3.5	0.58	0.36	0.39
Fewshot + Instruction + Definition	GPT-4	0.73	0.63	**0.66**
	Mixtral	0.54	0.34	0.30

Table 3: LLM classifiers performance comparison

the founding fathers of America said "All men are created equal." They referred to under the laws of the land. NOT equal in physical ability and mental capacity. IF all men were in fact created equal, thing such as affirmative action would not need to exist.' was tagged as Weak Favor instead of No Stance.

Additionally, the results show that the model have made errors in classifying 'Strong Against vs. Weak Against, suggesting the difficulty of detecting whether a post has a strong or weak position, i.e., *"Leftists are against equal opportunity by supporting AA and giving benefits to minorities, that is against of being liberal."* was tagged as Weak Against while our annotators labeled it as Strong Against.

These analysis suggest that while the LLM-Prompt combinations are proficient in many aspects, they may still face challenges in accurately identifying the nuances of stance, particularly when differentiating between varying degrees of agreement or opposition.

4.1 Temporal Analysis of Stance

We leveraged the best model-prompt pair (GPT-4 + (F+I+D)) to annotate the full dataset (N= 2,548 posts) from 23 selected subreddits. Table 1 (#Posts (All)) shows the total number of posts labeled in our data. We further analyzed stance of each subreddit. Table 2 displays the top three frequent stance categories within the top 15 subreddits. The results shows that these communities share a diverse range of positions. While majority of the subreddits have either taken a neutral position or engaged in questions and asking others' opinions as their most prevalent position toward affirmative action, a few subreddits have exhibited stances against AA, followed by a smaller share of posts that have showed skepticism to it.

Figure 1 shows the temporal shifts (using moving average technique with 2 year windows) in the number of posts with different stances over the time period of our dataset. As the figure suggests, while Question was the most frequent type until 2014 and witnessed a peak around 2013, No Stance and Strong Against started to emerge as the most prevalent stance for AA in the discourse around 2014 and experienced a noticeable peak in 2018 and 2019. Two positions of support for AA (Strong Favor and Weak Favor) also started to rise around 2017 and experiences their peak in 2019.

4.2 Prevalent Key-phrases

Table 6 shows the most frequent key-phrases extracted from the posts with each stance label. Across the stance categories, some of the most prevalent keyphrases include mentions of 'Race' and 'Racism', as well as some important social and cultural concepts such as 'Discrimination', 'Minorities', 'Diversity' and 'Privilege'. Another frequent theme among the keyphrases is 'Students,' 'Colleges' and 'Schools,' highlighting the importance of discussions around college admission and students experiences in the affirmative action discourse. We also observe a theme that mentions various racial, ethnic groups e.g. 'Asians', 'Blacks', 'whites', showcasing the involvement of different demographic groups in debates around AA. Some other keyphrases relate to certain political leanings e.g. 'Liberals', 'Conservatives', 'Libertarians' or role of the government and policy making.

5 Discussion

Affirmative Action is a complex issue with different stakeholders that has a wide range of cultural,

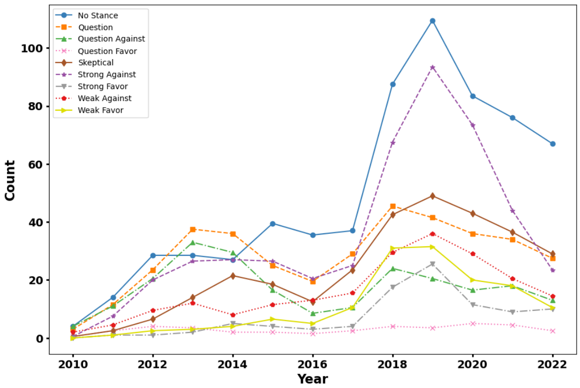

Figure 1: Temporal changes in number of posts with each stance over years

societal, and political implications. Exploring the online discourse on AA can enable us to shed light on how the public views various dimensions of this issue.

After annotating 400 posts and developing a ground truth dataset for our models, we used LLM-based models for stance classification. After experimenting with three LLMs and using different versions of our carefully curated prompts, we achieved the F1 score of 0.66, a competitive score compared to the state-of-the-art models (AlDayel and Magdy, 2021; Cruickshank and Ng, 2023). This promising result proves our hypothesis that the proposed expansion of stance classes can enhance our ability to detect a larger variation of stances. We demonstrate that adding to the prompts' complexity by providing clear instructions, definitions, and examples can boost LLM's performance in detecting the nuances of stance. It is important to note, however, that our study's goal is not to achieve the highest classification metrics. Instead, our motivation is to expand our understanding of stance, and offer a new way of thinking about its analysis.

Our error analysis shows the most common incidents of misclassification by the LLM classifier. This includes errors made when the model mistakes a neutral position with a skeptical or weak position, as well as mistakes in distinguishing strong from weak versions of a stance. This suggests that despite the overall success of our model, detecting the differences between these positions can still be a challenging task and needs further improvement. This could be linked to some of the known reasons for shortcomings of stance detection such as insufficiency of annotated data for the context of study or the differences between the language used for training LLMs and the language used in social media.

The results of our subreddit-level stance analysis highlight the prevalence of neutral (No Stance) and seeking other users' opinions (Question) on many of the popular Reddit communities, confirming the tendency of many users on this platform to discuss important issues with others and exchange ideas. This can also be suggested by the names of these subreddits such as *AskReddit*, *AskSocialScience* and *askaconservative*. For example, *r/AskReddit*'s goal based on its about page is for its subscribers "to ask and answer thought-provoking questions" and *r/asksocialscience*'s aim is "to provide great answers to social science questions, based on solid theory, practice, and research."

On the other hand, several prominent sociopolitical subreddits such as *unpopularopinion*, *changemyview*, and *TrueUnpopularOpinion* have expressed a strong stance against affirmative action. This observation can perhaps be explained by how these communities are named, encouraging their users to express unpopular, unconventional, and sometimes extreme ideas. *r/unpopularopinion* is a subreddit with more open discussions on controversial topics and *r/changemyview* is "A place to post an opinion you accept may be flawed, in an effort to understand other perspectives on the issue."

The temporal shifts in the frequency of different stances over the timeline of our dataset indicate that until 2014, the majority of posts are inquisitive in nature (with more Questions). We observed an increase in Question stance, with a peak in 2013. This can be linked to the Supreme Court's decision in the case of Fisher vs. University of Texas at Austin (Purdy, 2014). However, around 2017, No Stance and Strong Against began to dominate the Reddit discourse on AA and experienced a dramatic increase in the number of posts in 2019. This phenomenon may be related to the reversal of certain affirmative action guidelines by the Trump administration (Times, 2018). For example, a post from 2018 in our dataset mentioned: *"Is Trump pro-asian? Don't get me wrong, I've never liked him. But Think about the effects of his policies End to Affirmative Action. Some people say it's only going to benefit white people, but whatever you say an end to AA is a net benefit to Asian Americans."*

Another interesting observation is that the number of posts with favorable views towards AA remains persistently low, suggesting the largely unpopular attitude of Reddit users toward AA policies.

This result conflicts with the findings of several surveys done on the U.S. population that suggest the majority of Americans tend to favor AA in colleges, despite opposing explicit racial preferences in college admissions (Institute, 2024; Pew Research Center, 2023). Research also demonstrates that public attitudes about AA depend on how people are asked about it or the specific context in which it is being discussed, such as in higher education or the workplace (Pew Research Center, 2023; Petts, 2022). These findings align with our observations when annotating the posts.

During data annotation, we noticed a significant number of posts showed implicit support for the core idea behind AA, which is increasing the representation of marginalized people and underrepresented communities in college and workplace and rectifying some of the historical injustices while disapproving of the current implementation of AA policies and the explicit use of race or gender as a factor in assessing applicants. For instance, many users suggested that AA policies should be based on income level and economic circumstances, instead of race. i.e., *"Is AA on the basis of race more important than AA on the basis of economic circumstances? Is racial diversity more important than other forms of diversity? Is it somehow more important to achieve racial representation than economic representation? Why?"*, and *"I understand that the target is diversity and combating institutional racism, but on the basis of statistics, that could be achieved purely through looking at families' incomes. Blacks come from statistically poorer families and neighborhoods, so naturally, when you give an advantage to those with lower incomes, you're also significantly advantaging the vast majority of them and achieving diversity."* are two examples of such narratives.

The unfavorable view of Reddit users toward AA and its contrast with the overall support of the U.S. population for these policies can imply that the users on this platform tend to share ideas that are relatively controversial and different from the average population. In addition, the anonymity of users on platforms such as Reddit can exacerbate such viewpoints as users can express various forms of offensive, harmful, and sometimes hateful views without any serious consequences. This phenomenon can raise serious concerns over the state of online discourse and its impact on the epistemic welfare of society.

Due to the constraints of pushshift API, our dataset does not include the posts around the time the Supreme Court ruled against affirmative action in June 2023. This landmark decision sparked a new wave of debates around AA and similar policies. Therefore, we expect to witness another significant rise in the number of Reddit posts prior to and following this consequential decision.

The extracted keyphrases from the posts reveal several themes. Notions of 'Race' and 'Racism' and concepts of 'Discrimination', 'Minorities', 'Diversity' and 'Privilege' appear consistently across posts with different stances, illustrating the complicated nature of this topic and its various implications in society. On the other hand, the repetitive mentions of various demographic groups and political affiliations showcase multiple dimensions and stakeholders involved in AA discourse.

6 Conclusion and Future Work

Stance detection is a critical tool in better capturing diverse public views toward a range of controversial issues such as affirmative action. We demonstrated the usefulness and applicability of a new set of stance categories in successfully identifying and detecting the public stance toward affirmative action. Particularly, breaking down 'Favor' and 'Against' stances into their strong and weak versions, and adding various types of 'Question' and 'Skeptical' provide a more nuanced understanding of a diverse discourse. We collected over 2500 posts from 23 Reddit communities that concentrate on social, cultural, political discussions, and showed the prevalence of neutrality, questioning and strong opposition toward affirmative action among the users.

Future work includes experimenting the applicability of these new stances into other controversial domains with conflicting views such as abortion, feminism, climate change etc. Furthermore, we will investigate whether and to what extent multiple stances can co-occur in a post. While we used single stance annotation in this work, exploring the availability of multiple stances will further strengthen and enhance our nuanced understanding of the discourse. Lastly, to explore the variations in the affirmative action discourse, it is recommended to compare the language used in social media posts with how news outlets and mass media frame the issue.

7 Limitations

This work includes several limitations. First, despite our best efforts to include a wide range of opinions, the dataset we collected from 24 subreddits does not cover the entirety of Affirmative Action discourse, as we manually filtered out Reddit communities that were irrelevant to societal, cultural, and political discussions. Additionally, we acknowledge that Reddit users do not represent the entire society since these users tend to be from a young, tech-savvy, and college-educated background. Such drawbacks impede our ability to generalize the findings of this study on affirmative action to a broader population. Second, we only explored the potential of the 9 new stance categories in improving our understanding of Affirmative Action policies, not having tested our approach on other datasets from domains that draw public attention. Finally, in this work, we only used one stance label per post that captures the author's position most succinctly. However, in several occasions, more than one label may be applicable to a post, as also noted by our annotators. This was particularly the case in a number of posts labeld as 'Skeptical,' and 'Weak Against'. We included a note in our final annotated dataset for this situation.

8 Ethics Statement

Our data is primarily coming from English-speaking populations on one specific social media platform, which may not be generalizable to other linguistic or cultural contexts. All data was publicly available at the time of collection, and no direct interaction occurred between researchers and users. We adhere to strict data protection measures and have slightly altered the quotes to preserve anonymity and post integrity. We are using LLMs that may perpetuate biased ideologies and viewpoints. Our model may occasionally overgeneralize the stance, struggling to accurately capture the intensity of the speaker's position and the underlying intentions of the posts. This tendency to overgeneralize could stem from a variety of factors, including the inherent limitations of the training data or the model's architectural biases. Addressing these challenges will require further refinement of the model's training process and perhaps more nuanced prompt engineering to better differentiate between subtle variations in stance and intensity.

References

Abeer Aldayel and Walid Magdy. 2019. Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Fien Andries, Katharina Meissl, Clarissa de Vries, Kurt Feyaerts, Bert Oben, Paul Sambre, Myriam Vermeerbergen, and Geert Brône. 2023. Multimodal stance-taking in interaction—a systematic literature review. *Frontiers in Communication*, 8:1187977.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Andreas Bengtson. 2024. Affirmative action in the political domain. *Political Studies*, 72(1):158–176.

Karlyn H Bowman and Eleanor O'Neil. 2016. *Public opinion on affirmative action*. American Enterprise Institute for Public Policy Research.

Ryan Boyd, Steven Wilson, James Pennebaker, Michal Kosinski, David Stillwell, and Rada Mihalcea. 2015. Values in words: Using language to evaluate and understand personal values. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 31–40.

Steven M Cahn. 2013. *The affirmative action debate*. Routledge.

Erwin Chemerinsky. 1996. Making sense of the affirmative action debate. *Ohio Northern University Law Review*, 22:1159.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. *arXiv preprint arXiv:1906.03538*.

Major G Coleman. 1999. Merit, cost, and the affirmative action policy debate. *The Review of Black Political Economy*, 27(1):99–127.

E Couper-Kuhlen and M Selting. 2017. Stance and footing. *Interactional linguistics: Study language in social interaction*.

Kimberlé Crenshaw, Neil Gotanda, Gary Peller, and Kendall Thomas. 1995. *Critical race theory: The key writings that formed the movement*. The New Press.

Iain J Cruickshank and Lynnette Hui Xian Ng. 2023. Use of large language models for stance classification. *arXiv preprint arXiv:2309.13734*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dinesh D'Souza and Christopher Edley Jr. 1996. Affirmative action debate: Should race-based affirmative action be abandoned as a national policy. *Alb. L. Rev.*, 60:425.

Stanley Fish. 2000. The nifty nine arguments against affirmative action in higher education. *The journal of blacks in higher education*, (27):79–81.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Pentti Haddington. 2006. Identity and stance taking in news interviews. *Vestergaard, T., Strunck, J., & Lassen, I.(2006). Mediating Ideology in Text and Image: Ten Critical Studies. Amsterdam: John Benjamins Publishing Co*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. *arXiv preprint arXiv:2104.07467*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Cato Institute. 2024. Americans say they are for affirmative action yet oppose racial preferences in college admissions. https://www.cato.org/blog/americans-say-they-affirmative-action-yet-oppose-racial-preferences-college-admissions. Accessed: 2024-05-31.

Alexandra Jaffe. 2009. *Stance: sociolinguistic perspectives*. Oup Usa.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. CommunityLM: Probing partisan worldviews from language models. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. 2023. From values to opinions: Predicting human behaviors and stances using value-injected large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15539–15559, Singapore. Association for Computational Linguistics.

Randall Kennedy. 1985. Persuasion and distrust: A comment on the affirmative action debate. *Harv. L. Rev.*, 99:1327.

Dilek Kucuk and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6299–6305, Hong Kong, China. Association for Computational Linguistics.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. *arXiv preprint arXiv:2402.14298*.

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

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Amy L Petts. 2022. Attitudes about affirmative action in higher education admissions. *The Sociological Quarterly*, 63(4):711–732.

Pew Research Center. 2023. Americans and affirmative action: How the public sees the consideration of race in college admissions, hiring. https://www.pewresearch.org/short-reads/2023/06/16/americans-and-affirmative-action-how-the-public-sees-the-consideration-of-race-in-college-admissions-hiring/. Accessed: 2024-06-15.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2019. STANCY: Stance classification based on consistency cues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6413–6418, Hong Kong, China. Association for Computational Linguistics.

R Lawrence Purdy. 2014. Fisher v. university of texas at austin: Grutter (not) revisited. *Mo. L. REv.*, 79:1.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Rezvaneh Rezapour, Ly Dinh, and Jana Diesner. 2021. Incorporating the measurement of moral foundations theory into analyzing stances on controversial topics. In *Proceedings of the 32nd ACM conference on hypertext and social media*, pages 177–188.

Rezvaneh Rezapour, Saumil H Shah, and Jana Diesner. 2019. Enhancing the measurement of social effects by capturing morality. In *Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 35–45.

Charles H Rowell and Patricia J Williams. 1996. An interview with patricia j. williams. *Callaloo*, 19(4):823–834.

Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*.

Eyal Sagi and Morteza Dehghani. 2014. Measuring moral rhetoric in text. *Social science computer review*, 32(2):132–144.

Dan Goldwasser Shamik Roy, Nishanth Sridhar Nakshatri. 2022. Towards few-shot identification of morality frames using in-context learning. *Fifth Workshop on Natural Language Processing and Computational Social Science*.

Arushi Sharma, Abhibha Gupta, and Maneesh Bilalpur. 2023. Argumentative stance prediction: An exploratory study on multimodality and few-shot learning. In *Proceedings of the 10th Workshop on Argument Mining*, pages 167–174, Singapore. Association for Computational Linguistics.

Vasiliki Simaki, Carita Paradis, Maria Skeppstedt, Magnus Sahlgren, Kostiantyn Kucher, and Andreas Kerren. 2020. Annotating speaker stance in discourse: the brexit blog corpus. *Corpus Linguistics and Linguistic Theory*, 16(2):215–248.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77.

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. Association for Computational Linguistics.

Supreme Court of the United States. 2023. Students for fair admissions, inc. v. president and fellows of harvard college. `https://www.supremecourt.gov/`. Legal case where the Supreme Court held that race-based affirmative action programs in college admissions processes violate the Equal Protection Clause of the Fourteenth Amendment.

R. Roosevelt Thomas. 1990. From affirmative action to affirming diversity. *Harvard business review*, 68 2:107–17.

The New York Times. 2018. Trump officials reverse obama's policy on affirmative action in schools. `https://www.nytimes.com/2018/07/03/us/politics/trump-affirmative-action-race-schools.html`.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. 2022. Openstance: Real-world zero-shot stance detection. *arXiv preprint arXiv:2210.14299*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

A List of Subreddits

Table 4 shows the list of final subreddits selected for our study. Based on the objectives of these subreddits and close-reading of posts we found these subreddits to be the most relevant for the purpose of our study.

B Stance Categories and Definitions

Table 5 shows the final stance labels defined for our study, along with their definitions and an example of each.

C Keywords and Themse

Table 6 shows the extracted keyphrases for each stance cluster.

Subreddits
ApplyingToCollege, asianamerican, AsianMasculinity, askaconservative, AskALiberal, AskAnAmerican, asklatinamerica, AskReddit, AskSocialScience, changemyview, Destiny, Libertarian, neoliberal, NeutralPolitics, NoStupidQuestions, popularopinion, Residency, sociology, supremecourt, TooAfraidToAsk, TrueAskReddit, TrueUnpopularOpinion, unpopularopinion

Table 4: List of selected subreddits for analysis

D Prompts

Figure 2 shows the prompts used for zero and few-shot models for all LLMs. We used a combination of Instructions and/or Definitions as well as Examples for generating labels for each post.

Stance	Definition	Example
Strong Favor	The speaker strongly supports and defends affirmative action policies, discussing their necessity and benefits, often using emotional language in supporting AA.	Affirmative action is absolutely necessary to increase the diversity of minority groups in college and create a more equal society.
Strong Against	The speaker strongly opposes affirmative action policies, their harm and unfairness and discrimination, often using emotional language against AA.	There is no question that affirmative action is racist. Considering someone's race above merit for college admission is just wrong.
Weak Favor	The speaker supports affirmative action but with less intensity and certainty, maybe even critiquing some aspects of AA but showing support for it overall.	I think affirmative action, with its flaws, is still needed. We have not yet reached a point of true equality so the policies that support minority groups are still required in my opinion.
Weak Against	The speaker opposes affirmative action but with less intensity and certainty, maybe even supporting some aspects of AA but showing opposition to it overall.	As a Hispanic I dislike affirmative action on the basis of it being poor compensation for inadequate K-12 schooling. I'll preface this by saying I am a nerdy, progressive, and Hispanic college student, so this will probably anger people on the left and right.
Question	The speaker is neutral, asking a relevant question or seeking opinions of others about affirmative action without showing a personal stance.	What do you all think about affirmative action? I have some information about it but want to see what others think.
Question Favor	The speaker asks a question or seeks opinions of others about AA, with implicit support for affirmative action.	Which country (in the world) has done the most for its disadvantaged groups? Which country's affirmative action program has been the most successful?
Question Against	The speaker asks a question or seeks opinions of others about AA, with an implicit opposition to affirmative action.	Why don't college/job applications just remove the "What race are you?" and "What gender are you?" from the application? Wouldn't a race/gender blind application process, where applicants are judged only on grades/test scores, mean equality and fairness for all?
Skeptical	The speaker raises doubt or skepticism about the relevance, effectiveness or justification of affirmative action, or suggesting alternatives to change the current version of AA.	Affirmative action should be focused on funding better primary and secondary education in suburbs with poor minorities. The exams are held BEFORE they get into higher education, isn't the problem there? Affirmative action is a band aid on a bigger problem [...].
No Stance	The post does not mention affirmative action or AA is not the main topic discussed in the post, or the speaker remains fully impartial and neutral.	Asian Youtuber interviews Asian Harvard students about its affirmative action policies...and most of them are too scared to talk about it.

Table 5: Definition and example of stances

Stance	10 Most Frequent Keyphrases
Strong Against	Discrimination, Racism, Minorities, Race, Diversity, Colleges, Action Racist, Action Racism, Asians, Blacks
Strong Favor	Minorities, Racism, Discrimination, Privilege, Blacks, Males, College Admissions, Students, Minority Students, Racism Privilege
Weak Against	Racism, Minorities, Asians, Colleges, Race, Schools, Women, People, Governments, Students
Weak Favor	Racism, Admissions, Colleges, Discrimination, Minorities, Race, Women, Privilege, Asians, Schools
Question	Race, Colleges, Minorities, People, Schools, Students, Asians, Job, Applicants, Policy
Question Against	Racism, Minorities, Race, Colleges, People, Asians, Discrimination, Americans, Diversity, Schools
Question Favor	People, Colleges, Education, Blacks, Discrimination, Program, Minority Groups, Increase Diversity, Opportunities, Things
Skeptical	Minorities, Discrimination, Race, Diversity, Racism, Policies, Applicants, Education, Colleges, Women
No Stance	Racism, Colleges Admissions, Asians, Minorities, Liberals, Discrimination, Court, Libertarians, Government, Conservatives

Table 6: Most frequent key-phrases for each stance group

Instruction:
Read the following statement and determine its stance towards Affirmative Action. Respond with one of these labels: 'Strong Favor', 'Weak Favor', 'Strong Against', 'Weak Against', 'Question Favor', 'Question Against', 'Question', 'Skeptical', or 'No Stance'.…

Definition:
Affirmative Action Definition: Affirmative Action (AA) represents a set of policies and practices aimed at addressing historical and ongoing inequalities in employment, education, and other sectors by providing opportunities to historically marginalized groups. These measures are designed to promote diversity and rectify socio-economic disparities caused by past discrimination.
Definition of Stance: Stance is the expression of the speaker's standpoint and judgment toward a given target, usually a controversial cultural, social, or political topic such as affirmative action. The definition of labels:
Strong Favor: The speaker strongly supports and defends affirmative action policies, discussing their necessity and benefits, often using emotional language in supporting AA.
Weak Favor: The speaker supports affirmative action but with less intensity and certainty, maybe even critiquing some aspects of AA but showing support for it overall.
Strong Against: The speaker strongly opposes affirmative action policies, their harm and unfairness and discrimination, often using emotional language against AA.
Weak Against: The speaker opposes affirmative action but with less intensity and certainty, maybe even supporting some aspects of AA but showing opposition to it overall.
Question Favor: The speaker asks a question or seeks opinions of others about AA, with an implicit support for affirmative action.
Question Against: The speaker asks a question or seeks opinions of others about AA, with an implicit opposition to affirmative action.
Question: The speaker is neutral, asking a relevant question or seeking opinions of others about affirmative action without showing a personal stance.
Skeptical: The speaker raises doubt or skepticism about the relevance, effectiveness or justification of affirmative action, or suggesting alternatives to change the current version of AA.
No Stance: The post does not mention affirmative action or AA is not the main topic discussed in the post, or the speaker remains fully impartial and neutral. Also, posts that are very long and contain several paragraph discussing several issues but only briefly mention AA, for example in just one sentence, belong to this group since affirmative action is not the main idea in the post.

Example:
Post: Supreme court will decided next whether affirmative action should be continued or not
Label: No Stance
Post: Affirmative action should be based on income level and socioeconomic status, not race, to achieve its goal. America is divided based on class, not race.
Label: Skeptical
...

Task:
What is the stance of the following statement toward affirmative action?
Constraint:
Only return the stance using one of the mentioned labels , and no additional text.
Here is the statement:{post}

Figure 2: Prompts used for classification of stance