Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science and Media

Author or Argumentation: Exploring the Effect of Prior Beliefs and Personality Traits on Persuasiveness

Master's Thesis

Nikolay Kolyada
Born Apr 27, 1984 in Leningrad

Matriculation Number 116561

1. Referee: Prof. Dr. Benno Stein
2. Referee: Prof. Dr. Andreas Jakoby

Submission date: June 5, 2019

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, June 5, 2019

.....
Nikolay Kolyada

Abstract

This work aims to explore the influence of the mutual interests, prior-beliefs and personality traits of discussion participants on their persuasiveness, i.e. the success in changing the opponent's opinion in online discussion forums. To answer the question of whether mutual author beliefs and personality traits between two debaters are as important for persuasiveness as argumentation quality, we introduce several author–based features and compare them to text-based features such as grammatical, stylistic and lexical features on two persuasiveness tasks.

We use the corpus of the Reddit website as the most popular and vast online forum resource. Particularly, we explore the "Change My View" subreddit, a community on Reddit specifically devoted to detailed online discussions of various topics, with strict moderation rules and an extensive number of discussions. It implements a special set of "Delta"-rules that help explicitly label persuasive arguments. We use the data from "Change My View" to prepare an input dataset for the main task of predicting the winner of a discussion from a pair of selected participants. As an additional task, we explore the influence of personality traits on the malleability of the original poster's point of view.

Our model of authors relies on two bases: (1) author interests and beliefs inferred from the history of their previous activity in different Reddit communities and (2) author psychological traits theory models such as the Big Five personality traits model.

We experiment with various author features representing user interests and prior beliefs and observe a noticeable improvement over a baseline of argumentation-level features on persuasiveness detection, that is, author-level features achieve significant improvement of accuracy in the two studied tasks.

Contents

Chapter 1

Introduction

The art of persuasion is arguably one of the most important communication skills. Changing someone's opinion is a common goal in many different settings from private discussions to political propaganda. However, the underlying mechanisms of persuasion are complex and very difficult to study [Petty, 2018].

Nowadays, with ubiquitous media technologies, persuasive strategies are used by the press, governments, advertising companies and other members of the media ecosystem. Especially the Internet and social media are ways of communication with the audience that provide vast opportunities for persuasive mass communication [Fogg, 2008]. That makes it worthwhile to explore the mechanisms of persuasion, and to find out what influences persuasiveness the most.

This thesis at hand studies how the similarity of users interests, beliefs, and psychological traits can influence their persuasiveness in online debates. In particular, we hypothesize that there might be a correlation between the similarity of two participants on a discussion and the likelihood of one of them to change her current points of view.

Several studies in the areas of psychology and computational argumentation target the persuasiveness or convincingness of arguments from the points of view of source credibility [Pornpitakpan, 2004, Sternthal et al., 1978, Umeogu, 2012], types of evidence [Hornikx, 2005], properties of arguments [Habernal and Gurevych, 2016a,b], linguistic and interaction patterns [Tan et al., 2016] and prior beliefs and personality traits of the discussion participants [Durmus and Cardie, 2018, Lukin et al., 2017]. Regarding the former line of research, the way of modeling prior beliefs and personality traits is either by deriving certain information from the users accounts or by asking the users to have psychological tests. Up to our knowledge, no prior work on analyzing persuasiveness

that considers personal-level features where such features are constructed *automatically* from the history of the discussion participants' activity.

In this work, we focus on aspects of persuasion related to the discussion participants rather than features of argumentation, while we compare the performance of proposed features with argument-level features.

Our research questions are formulated as follows:

- How to model the similarity of people's interests, beliefs, and personality traits?
- How to operationalize the developed model in terms of identifying the the model's attributes automatically?
- How to demonstrate the impact of the similarity of people's interests, beliefs, and personality traits on their persuasiveness, if any?

Analyzing the user content generated from the complete set of Reddit communities, first, we propose 39 types of author-level features including a set of person's interests, her prior-beliefs (as for example polarity of opinion about controversial topics) and psychological characteristics of person defined by the language he or she uses across different Reddit forums. Then, we compute the similarity of authors based on the created author-level features-vectors. For purposes of evaluation two datasets were extracted from the Reddit's community called "Change My View" [1] (CMV). The CMV implements a special awarding mechanism for discussion participants called "Delta system" which helps to automatically mark persuasive replies.

We conduct two experiments in order to evaluate author-level features: (1) we predict the most argumentative reply out of a pair of the comments from the same discussion and (2) classify the opinion expressed by the original poster as resistant or malleable. We hypothesise that the author similarity influence on persuasion success and assume that a larger amount of data about an author's past activity on Reddit would improve prediction accuracy. As it will be described later, it turns not to be the case in general. Our results show a noticeable improvement over a strong baseline in the case when textual features are used together with author-level features. For the task of classification of malleability the opinion, we report significant improvement of the prediction accuracy over the state-of-the-art. In addition, we analyse of the correlation

[1] `https://www.reddit.com/r/changemyview/`

between personality traits and opinion malleability and describe several interesting findings which comply with the definitions of Big Five personality traits.

We experimented with various features representing users and observed a noticeable improvement over baselines using argumentation-level features only. Therefore, we believe that the main contributions of the present work are: (1) the construction of different types of author-level features representing interests, prior beliefs and personality traits; (2) creating a large-scale dataset of pair of replies and original posts for evaluation of the proposed features; and (3) an improvement in the prediction accuracy for both tasks and insights into the interconnection between author-level features and persuasion success. We believe the ideas of the proposed approaches can be generalized and used in areas like marketing or advertising.

The remaining chapters are organised as follows: in the second chapter, we describe the corpora; lay out the background ideas for the scope of this work and give an overview of related work. The third chapter is devoted exclusively to the discussion of the steps taken during the work, including data pre-processing, feature engineering and chosen methods for approaching the research questions. After that, we evaluate the created models, perform analysis and compare the results. Finally, we summarise the experiments and discuss future steps that we believe are worth pursuing within the topic.

Chapter 2

Background

In this chapter, we describe the datasets and give an overview of the approaches used throughout the experiments. We also summarize related work.

In order to answer the research questions, we had to find suitable datasets. We consider the following requirements for our dataset: it should contain discussions in machine-readable form, the number of topics should be as broad as possible (not highly specialized forums), and contain a clear marker for discussions where the original point of view of a participant changes. It should contain additional data about members of discussions so it would be possible to model these users. The well-known online discussion forum Reddit and its community "Change My View" fits these requirements. Therefore it has become the main source for these experiments.

To approach the problem of user modelling, we try to represent user interests and beliefs using information about their previous activity on Reddit. We use an approach similar to the way of modelling users in the widely used Recommender systems[1]. We represent users as feature vectors of their posts in different subreddit categories and then compute similarities and explore how these similarities influence persuasion success.

We apply several typing approaches to model users: utilizing the topic category tree of communities to represent areas of interests of Reddit users and using post frequencies and content as input for author features. From the content of user posts, we extract named entities and context sentiment they were mentioned in. Also, we use the text from all user posts to construct personality traits features.

[1] `https://en.wikipedia.org/wiki/Recommender_system`

2.1 Corpora

The main corpora used for feature construction and the experiments is a Reddit crawl version of March 2018. In addition, Wikipedia and Wikidata datasets are used at some of the feature preparation steps.

2.1.1 Reddit

Reddit[2] is a massive news aggregation, user-generated content and discussion website founded in the United States of America and it is one of the most popular forums on the Internet. It was founded in 2005 and as of now is the 21st world top visited website in the world with 542 million monthly visitors according to Alexa.com. Registered users can post any kind of content including text, images, videos and links.

All posts are organised into user-managed groups or communities called "subreddits" covering numerous different topics such as politics, science, books, cinema, music, entertainment and others. Reddit uses a rating system for posts based on users' up-/down-votes; higher rated posts appear closer to the top of a subreddit's front page.

As of now, there are over 138,000 active subreddits with over 48 million unique users. The version of Reddit crawl used for this study covers the period from the foundation of the website in the middle of 2005 till the end of September 2017. It contains posts from almost 900,000 subreddits, around 540 million submissions and over 3.5 billion comments. We use the total of this data as an input for the user modelling tasks described in the following chapter 3.1. Some Reddit dataset statistics are shown in Table 2.1.

Dataset timespan	23.06.2005 – 30.09.2017
# subreddits	887 655
# submissions	540 034 904
# comments	3 662 675 324
# users	42 369 105
# users ever posted to CMV	174 873
# pairs	210 774

Table 2.1: Reddit crawl statistics.

[2] `https://www.reddit.com/`

Snoopsnoo

Snoopsnoo[3] is a third-party analytical tool and website which provides insights into the activity of Reddit communities. It indexes changes on Reddit on a daily basis and provides several tools to explore Reddit communities' as well as Reddit users' activity.

top category	# subcategories	# subreddits
Adult and NSFW	0	3729
Architecture	0	41
Art	1	153
Business	6	179
Education	1	332
Entertainment	176	2371
Gaming	126	2707
General	8	2613
Hobbies and Interests	64	1601
Law	0	32
Lifestyle	68	3057
Locations	112	1935
Meta	0	388
Music	21	992
News and Politics	18	872
Science	16	400
Social Science and Humanities	6	198
Sports	59	684
Technology	49	1730
Travel	0	41
Other	0	1 810 014

Table 2.2: Top-level subreddit categories statistics.

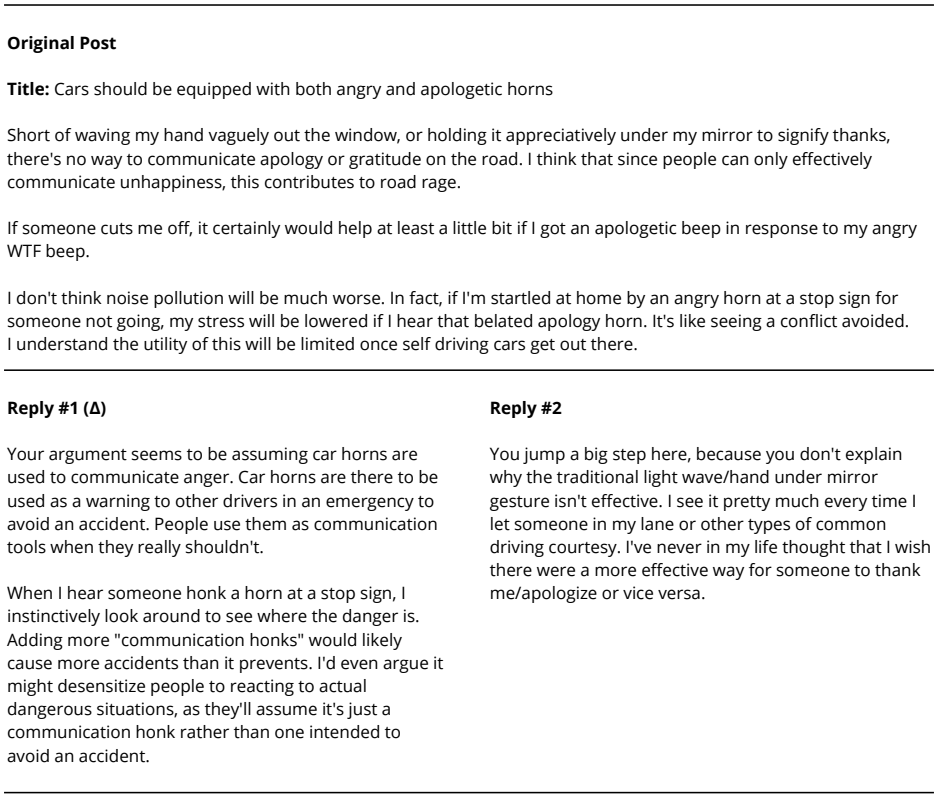
The categorisation of communities provided by Snoopsnoo is the most useful feature for our study. It classifies almost 100,000 of the most popular subreddits into 720 categories and subcategories. A three-level category system classifies subreddits by topic (i. e. Science –> Astronomy, Biology, ...) in parallel with other meaningful properties such as for example geographical positions (Locations –> Canada, Europe, ...). The statistics for top-level subreddit categories are shown in Table 2.2.

In this work, the whole of 720 categories grouped into three levels was crawled along with subreddits mapped to these categories. We deliberately excluded from the crawler the "Other" category which holds almost 1.8 million subreddits and at the step of feature engineering consider subreddits that are not assigned to any category as belonging to "Other".

[3] `http://snoopsnoo.com/`

2.1.2 ChangeMyView

While using the data from all subreddits for modelling authors, for construction of pairs dataset, we focus on the specific Reddit community called "Change My View" [4] (CMV). It was created at the beginning of 2013 and gained significant popularity during the following years. The version of Reddit dataset used in the work contains data with more than 174 000 users and over 60,000 separate discussions.



Original Post

Title: Cars should be equipped with both angry and apologetic horns

Short of waving my hand vaguely out the window, or holding it appreciatively under my mirror to signify thanks, there's no way to communicate apology or gratitude on the road. I think that since people can only effectively communicate unhappiness, this contributes to road rage.

If someone cuts me off, it certainly would help at least a little bit if I got an apologetic beep in response to my angry WTF beep.

I don't think noise pollution will be much worse. In fact, if I'm startled at home by an angry horn at a stop sign for someone not going, my stress will be lowered if I hear that belated apology horn. It's like seeing a conflict avoided. I understand the utility of this will be limited once self driving cars get out there.

Reply #1 (Δ)

Your argument seems to be assuming car horns are used to communicate anger. Car horns are there to be used as a warning to other drivers in an emergency to avoid an accident. People use them as communication tools when they really shouldn't.

When I hear someone honk a horn at a stop sign, I instinctively look around to see where the danger is. Adding more "communication honks" would likely cause more accidents than it prevents. I'd even argue it might desensitize people to reacting to actual dangerous situations, as they'll assume it's just a communication honk rather than one intended to avoid an accident.

Reply #2

You jump a big step here, because you don't explain why the traditional light wave/hand under mirror gesture isn't effective. I see it pretty much every time I let someone in my lane or other types of common driving courtesy. I've never in my life thought that I wish there were a more effective way for someone to thank me/apologize or vice versa.

Figure 2.1: Example of a CMV submission.

CMV is a subreddit devoted to organised discussions on diverse topics. It implements a special set of "Delta"-rules that help explicitly label persuasive arguments that is essential for our work. All other members of the discussion should directly reply to the comment that they consider persuasive with the

[4] `https://www.reddit.com/r/changemyview/`

special character "Δ" or the expression "!delta" which is a marker for the community's internal algorithm to explicitly mark the comment as containing persuasive argumentation. Basically, it is a discussion platform with its own publication and commenting rules. The general structure is similar to other online communities running on Reddit's engine: there is a front-page which is a board with posts sorted by up and down votes and post date/time.

In CMV a separate post, just as in any other subreddit, is called submission. It represents a topic that the original poster (OP) offers to the community members to discuss. It is an invitation for others to challenge the original opinion and provide argumentation to persuade the OP to change his or her mind. The Figure 2.1 shows an example of the submission along with two user comments from the discussion: one which was awarded a Δ from the original poster and the other which did not manage to persuade her.

ChangeMyView Rules

Following the publication rules, author of a submission should clearly state her stance or point of view on a certain problem and provide detailed reasoning to support of claims: "Explain the reasoning behind your view, not just what that view is (500+ characters required)." [5].

All replies to the original post are comments by a participant of the discussion. Top-level comments are direct responses to submission and considered as "head-arguments", therefore should follow similar to original post rules. All top-level replies should challenge OP: "Direct responses to a CMV post must challenge at least one aspect of OP's stated view (however minor), unless they are asking a clarifying question."[6].

Original poster cannot create top-level comments for any reason. In case there is need to clarify the view the original post should be edited.

There are several other rules that users of CMV should follow, such as moderation of rude/hostile comments, ads ban and others that increase the overall quality of the discussions but are less important for the experiments.

Delta System

The community implements a special awarding mechanism for discussion participants called "Delta system" that is truly valuable for us.

[5] https://www.reddit.com/r/changemyview/wiki/rules#wiki_submission_rules
[6] https://www.reddit.com/r/changemyview/wiki/rules#wiki_comment_rules

The Original poster and all other members of the discussion should directly reply to the comment that shifts their point of view to any degree and within their reply type the special character "Δ" or the expression "!delta". It is a marker for the community's bot called the DeltaBot, which is monitoring all new posts, to explicitly mark the comment as containing persuasive argumentation by awarding its author and the comment itself with the Delta sign. The amount of Deltas given is accumulated and showed in user profiles and next to a users' nickname across all discussions. It serves as an indicator of user experience and motivation for the members of the community.

The unique Delta system provides us with the opportunity to process all the data in CMV automatically and prepare an extensive and at the same time quite clean and reliable dataset for the exploration of different factors that have an influence on persuasion success.

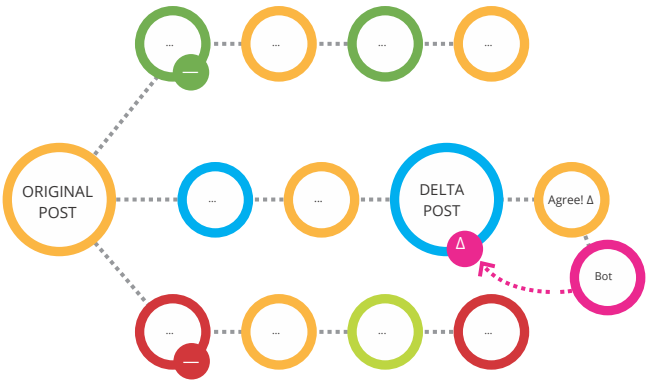


Figure 2.2: Example of a discussion tree of a CMV submission with three branches and one Delta-awarded comment.

The figure 2.2 shows an example of a submission structure with three branches of discussion (subtrees), the discussion flow goes from the original post on the left to later replies on the right respectively. The two subtrees with top-level replies marked with the "–" sign are branches which were not considered by the OP as persuasive. The posts are colour-coded by author, yellow posts are posts by the OP, blue post by a challenger which has managed to persuade OP and receive a Delta. You can see the most common patterns of CMV discussions: back-and-forth argumentation; comments never answered

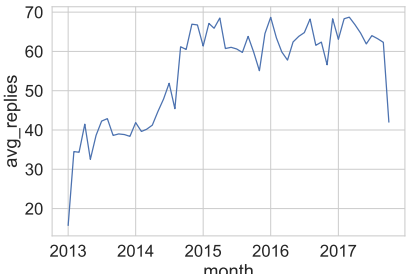
by OP; comments that did not succeed to persuade OP; replies were awarded Delta and Delta Bot automatic reply.

Dataset Analysis

We perform some initial analysis of the ChangeMyView dataset, gather general statistics and explore how interaction dynamics and some user properties influence persuasion process. We observe similar interaction patterns as Tan et al. [2016], albeit our dataset is covering three times longer period.

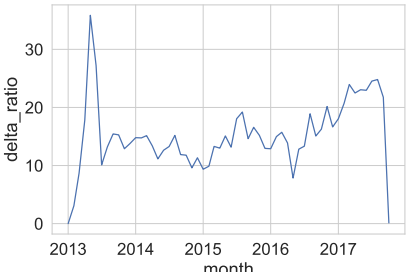
The Figure 2.3 displays how the number of posts, average number of replies/participants and percentage of Delta-awarded posts were changing over the lifetime of the community.

(a) Number of submissions per month



(b) Average number of replies per submission

(c) Average number of authors per submission



(d) Average delta percentage

Figure 2.3: Monthly activity in the dataset. The delta percentage is the fraction of discussions with Delta-awarded replies.

WinArgs dataset

As we employ certain approaches to the computation of features and construction of evaluation datasets as Tan et al. [2016] we make use of the dataset provided with the paper (WinArgs dataset) in our experiments in order to compare the results. Their version of CMV covers the period from the creation of the subreddit in January 2013 till September 2015. Table 2.3 provides general statistics of the WinArgs dataset.

# discussion trees	# nodes	# OPs	# uniq. authors
20,626	1,260,266	14,174	86,888

Table 2.3: Discussion trees dataset statistics.

2.2 Persuasion and Personality

The first attempts to provide the ground ideas of how persuasion works date back to 4th century BC to Aristotle and his "Rhetoric" tractate where he formulates key components of persuasion or "modes of persuasion" as the following three: Ethos, Logos and Pathos. [Braet, 1992]

While Logos according to "Rhetoric", at the first place specifies the quality of argumentation itself appealing to logic and supporting speaker's claims, Pathos and Ethos are related to characteristics of the speaker. Ethos implies an appeal to the authority of the speaker and indicates her credibility or character from the point of view of the audience. Pathos as an appeal to audience's emotions means the use of emotional tone, metaphors and passionate delivery to invoke sympathy from an audience.

Argumentation is the essence of the process of reasoning, the ability to make sense of things, apply logic, make claims and provide support for them. The field of logic studies the ways humans apply reasoning through arguments. The argument consists of a sequence of statements including claim and support or premises used by the arguer to convince the claim, they can include evidence and motivational appeals.

2.2.1 Personality Traits Theory

Personality traits theory is a field in psychology focusing on the study of human personality. It proposes approaches to the measurement of *traits* – behavioural, intellectual and emotional patterns. Personality traits can be seen as aspects of personality that form character, are consistent over situations and differ across individuals.

Psychologists measure personality through objective tests (self-reports) and projective measures. Objective tests rely on one's personal responses structured in a way that an individual has a limited set of options and are relatively free of bias. Some of the more widely used personality measures are the Myers-Briggs Type Indicator, Neo Pi-R, MMPI, and 16PF. Projective measures, on the other hand, origin in psychoanalytic theories of personality and involve using ambiguous stimuli to reveal the individual's hidden emotions and conflicts. [Coaley, 2010]

The one of most widely accepted of personality traits theories are 16PF and Big Five. Formulated by Raymond Cattell, an early proponent of using factor analytic methods "16 personality factor model" and 16PF Questionnaire

gained tremendous recognition across the field. Cattell's "source" traits are precursors of currently very popular "Big Five" (five-factor) model of personality. The initial model was proposed by Ernest Tupes and Raymond Christal in 1961 [Tupes and Christal, 1992] and later was advanced by Goldberg [1990]. The Big Five model is the result of factor analysis of intercorrelations of 16 primary trait measures proposed by Cattell. The five factors are include: *openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism* (OCEAN, CANOE).

The Big Five personality traits are defined as following[7]: "

- *Openness to experience* is one of the domains which are used to describe human personality in the Five Factor Model. Openness involves five facets, or dimensions, including active imagination, aesthetic sensitivity, attentiveness to inner feelings, preference for variety, and intellectual curiosity.
- *Conscientiousness* is the personality trait of being careful, or diligent. Conscientiousness implies a desire to do a task well, and to take obligations to others seriously. Conscientious people tend to be efficient and organized as opposed to easy-going and disorderly.
- *Extraversion* is the state of primarily obtaining gratification from outside oneself. Extraverts tend to enjoy human interactions and to be enthusiastic, talkative, assertive, and gregarious.
- *Agreeableness* is a personality trait manifesting itself in individual behavioural characteristics that are perceived as kind, sympathetic, cooperative, warm, and considerate.
- *Neuroticism* is one of the Big Five higher-order personality traits in the study of psychology. Individuals who score high on neuroticism are more likely than average to be moody and to experience such feelings as anxiety, worry, fear, anger, frustration, envy, jealousy, guilt, depressed mood, and loneliness.

"

Several studies explored associations between personality and language use in various different contexts. Results confirmed previous work on personality and identified predicted correlations between the psychological dimensions like Extraversion and Neuroticism and usage of words from different positive and negative emotion categories [Pennebaker et al., 2001].

2.3 Foundations of Computational Linguistics

Entity Linking is a task of recognizing and determining the identity of entities mentioned in the text. Also, known as Named Entity Recognition and

[7] `https://en.wikipedia.org/wiki/Big_Five_personality_traits`

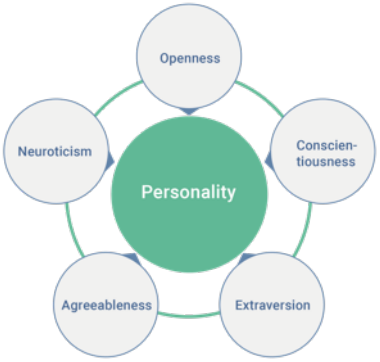


Figure 2.4: Five main personality traits according Big Five theory by Anna Tunikova for peats.de and wikipedia [CC BY 4.0 (https://creativecommons.org/licenses/by/4.0), `https://en.wikipedia.org/wiki/Big_Five_personality_traits`]

Disambiguation techniques let obtain mapping from entities mentioned in a text to a Knowledge base like Wikidata [8], DBpedia [9], YAGO [10] or Freebase [11]. Bridging Web data with knowledge bases is truly beneficial for annotating the raw and usually noisy data on the Web and contributes to the vision of Semantic Web [Berners-Lee et al., 2001]. The entity linking task is difficult due to the ambiguity of the meaning of the words and names and that named entities may have multiple forms such as full name, aliases and abbreviations.

Many approaches to entity linking exist. The state-of-the-art Raiman and Raiman [2018] reach impressive accuracy of $\sim$ 94.88% applying the solution on standard datasets (i.e. WikiDisamb30, CoNLL (YAGO), TAC KBP 2010). They construct a special type system and use it to constrain the outputs of a neural network to respect the symbolic structure.

Not focusing in this work on entity linking task itself, we used simple Link Probability approach [Milne and Witten, 2008] exploiting Wikipedia statistics and mappings between Wikipedia articles. Additionally, we integrate several Wikidata properties such as "instance of" and others into the model as classifying entity features to certain types such as persons or organisations.

Sentiment analysis or Opinion mining is a set of methods in the area of Computational Linguistics focused on automatic extraction of affective state

[8] `http://wikidata.org/`
[9] `http://dbpedia.org/`
[10] `http://yago-knowledge.org/`
[11] `http://www.freebase.com/`

and subjective information from written or spoken the language, classifying the polarity of the language used. It is useful to a wide range of problems that are of interest to researchers of such fields as computational linguistics, sociology, marketing and advertising, psychology, economics, and political science.

Existing approaches to sentiment analysis can be grouped into three main categories: knowledge-based techniques, statistical methods, and combined approaches. Knowledge-based approaches employ predefined dictionaries of classified unambiguous affect words such as happy, sad, afraid, and bored to classify the text. Statistical methods imply use of machine learning techniques such as latent semantic analysis, support vector machines, bag-of-words and deep learning. Hybrid approaches make use of machine learning methods as well as knowledge representation approaches like knowledge bases, ontologies and semantic networks in order to extract semantics out of the text. [Liu and Zhang, 2012]

For the tasks, we use one of the tools implementing the knowledge-based approach. "VADER (for Valence Aware Dictionary for sEntiment Reasoning) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media."[12] VADER uses a combination of a sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative. [Hutto and Gilbert, 2014]

LIWC – Linguistic Inquiry and Word Count

LIWC[13] (pronounced "Luke") is a computerized text analysis tool based on word-count approach. It outputs the percentage of words falling into over 80 linguistic, psychological, and topical categories. The LIWC2015 dictionary composed of almost 6400 words. It is a result of extensive research establishing correlations between linguistic patterns and personality or psychological state. The online version of LIWC API provides some additional personality traits features. [Tausczik and Pennebaker, 2010a]

LIWC has been widely used in the social media domain [Yarkoni, 2010] and due to its straightforward dictionary format, word lists can be easily understood, and extended if necessary.

[12] `https://github.com/cjhutto/vaderSentiment`
[13] `http://liwc.wpengine.com/`

2.4 Related Work

We utilize the same procedure and the experimental setup from Tan et al. [2016] to compute argumentation features for our dataset. We apply a similar dataset split approach and implement several most powerful features from the paper to compare the performance of author features. Since in Tan et al. explore the same dataset of ChangeMyView Reddit community and their version of crawl just covers smaller timespan, we expect to get the comparable results for the same features as they did. We consider their results as baseline-performance we try to improve.

In their work, Tan et al. [2016] conducted two experiments: prediction the winning argument in a pair of replies with similar argumentation and classify point of view expressed in the submission as resistant or malleable. We employ similar procedure to one in the paper to design our experiments, except that we explore additional author features along with the linguistic and "interplay" features.

Tan et al. [2016] find several interesting patterns of interaction dynamics, such as participant entry-order and degree of back-and-forth exchange and show that language factors play an essential role. Especially, the interplay between the language of the original poster and that of the challenger provides strong predictive features of persuasiveness. Investigating the problem of classification of original post's opinion as resistant or malleable they report that textual features of how the opinion is expressed carry predictive power.

The other work we referring to is Durmus and Cardie [2018] as one of the evidence for the hypotheses being true. Although Durmus and Cardie explore different corpus, their hypotheses share the same idea of the influence of prior-beliefs similarity on persuasion success in online debates. They introduce some author-level features and compare them with argument-level. They provide results showing that similarity of authors' prior-beliefs play a significant role in persuasion success.

In their work, Durmus and Cardie [2018] construct author features employing the data from user profiles as the Debate.org portal provides additional information about the stance of a user towards several most important topics like Politics or Religion. Controlling for topics of the debates, they evaluate how user similarity is correlated with persuasion success and compare prediction performance with textual features only. They report significant improvement in the setup with user-based features.

In the current work, we try to combine these two approaches and model author interests and beliefs as well as main personality traits. In contrast with Durmus and Cardie paper, we try to engineer author beliefs/interest features from their raw text and not from filled manually by user profile information.

Chapter 3

Modelling Authors and Argumentation

In this chapter, we describe the approaches to design author-level and textual features and prepare datasets for our experiments of studying how user similarity influences persuasion success. We propose three classes of author features: *Interests*, *Beliefs* and *Personality traits* and process information about user posts from across all Reddit communities to construct them. As textual features, we compute some of the most powerful ones as suggested by Tan et al. [2016], which include argument-only features such as length of posts and other representing the interplay between the replies and the original post. Finally, we describe the approach used for the preparation of the experiments datasets.

3.1 Feature Engineering

We propose three classes of author features: *Interests*, *Beliefs* and *Personality traits*. For each of the classes we use data from the same Reddit corpora, but focus on different aspects of the data. To model user *interests*, information about user posts across all Reddit communities is used with the number of posts as a base feature. In order to represent *beliefs* of a user, we extract named entities out of user post texts along with the sentiment of the context they are mentioned in. The third class of features - *personality traits* of authors includes data provided by LIWC API (2.3). From the output of the API following types of features are collected: 85 psychological word-categories, Big Five (2.4) and 54 other personality traits.

The features representing interests and beliefs are divided into two "levels": the sparse "flat" features and "fine-grained" features, where each user is represented as a set of smaller feature vectors for each of subreddit cate-

gory individually. This way we expect to get a more specific representation of author interests within a separate topic. Altogether, three proposed feature classes contain over 39 groups of features and over 7 000 separate features if consider "fine-grained" features within categories.

As we will describe in detail in the next chapters, the author features are evaluated in the two following experiments:

- First task of predicting the winning argument (4.1) from a pair of similar argumentative comments. For these tasks, all proposed classes of features are used: interests, beliefs and personality traits.
- For the second task of predicting malleability of the original poster's point of view (4.2), we experiment only with author traits features.

3.1.1 Subreddit Topics as Indicators of Interests

We make an assumption, that activity of a user in different communities across Reddit reflects her interests. For example, if a certain Reddit user living in New York City has as such hobbies as photography and film making, it is likely that she at least once has posted to subreddits belonging to categories like Hobbies and Interests –> Photography, Technology –> Gadgets and probably Locations –> United States –> New York City if she is taking an active part in the life of local communities. We expect, that the more active user participates in discussions on a certain topic, the more explicit is her interest in the topic and more representative such data is.

To model the interests, information about user posts from all Reddit communities is considered. Raw post frequencies are used as basic features and later as input for more complex like "fine-grained" category features. We also apply following techniques to compute feature values: normalized counts, one-hot encoding, TF-IDF and Principal Component Analysis[1].

Subreddit Feature Vectors

At this first step, we approach user modelling task in a simple way: we represent users as vectors of their post frequencies in different subreddits. Similarly to well-known Bag-Of-Words[2] approach, we treat authors as documents and subreddits as if they were terms in those documents. Although, it is not our

[1] https://en.wikipedia.org/wiki/Principal_component_analysis
[2] https://en.wikipedia.org/wiki/Bag-of-words_model

goal to search for most similar users, we need a way to measure the similarity of an arbitrary pair of authors. We collect all submissions and comments, group them by authors and transform into a vector representation. As a result, we get very sparse vectors with total 81 951 of "subreddit"-features.

To compute feature vectors we apply several approaches:

- raw count - sum of submission and comments (post frequency);
- one-hot representation of feature vectors for all subreddit/category/top-category types of feature-vectors;
- weighted counts with TF-IDF scheme.

For TF-IDF, we define Term Frequency as:

$$TF_{author,subreddit} = \frac{|Post_{author,subreddit}|}{|Post_{author}|}$$

where $P_{author,subreddit}$ are posts by the author to a certain subreddit and P_{author} are posts by author across all subreddits.

Interpreted as document frequency, "author frequency" is a number of authors posted to each of subreddits. We define IDF as:

$$IDF_{subreddit} = \log(\frac{|A|}{|A_{subreddit}|})$$

where A are all authors in the dataset and $A_{subreddit}$ are authors that have at least one post to the subreddit.

"Fine-grained" Category Features

As one of the ways to tackle the sparsity of subreddit-based features vectors, we introduce other class of features: using subreddit category tree2.1.1 and generalising user activity to 20 top-categories and 720 sub-categories, we construct "fine-grained" features. Each user, in this case, is represented as a set of smaller subreddit-vectors within each category and top-category. In this way, we obtain more precise, "fine-grained" feature vectors representing author interests within a certain topic.

The used category hierarchy classifies over 26 000 (excluding the 'Other' category) most popular subreddits into three-level category system of 720 categories representing different topics (ex. Science –> Astronomy, Biology, ...)

along with other meaningful properties such as geographical (Locations –> Canada, Europe, ...).

For all fine-grained features, several approaches to the calculation of weights are used. Along with raw counts of author posts we compute normalized counts, TF-IDF weighted and one-hot encoded vectors. In the case with top-level category features the cumulative post frequencies from all child subcategories are taken into account.

For TF-IDF weights categories are interpreted as terms instead of subreddits, as following:

$$TF_{author,category} = \frac{|Post_{author,category}|}{|Post_{author}|}$$

where $P_{author,category}$ are posts by the author to a certain category and P_{author} are posts by author across all categories.

Category document frequency are defined as:

$$IDF_{category} = \log(\frac{|A|}{|A_{category}|})$$

where A are all authors in the dataset and $A_{subreddit}$ are authors that have at least one post to the category.

3.1.2 Prior Beliefs: Extracting Author Stance

The other proposed class of author-level features is *prior-belief* features. The prior-beliefs suppose to represent the author's stance towards different targets. The target may be a person, an organization, geographical location, a product or any abstract concept containing in the Wikipedia. The targets are represented by named entities extracted from the author's posts. The method chosen for named entity extraction is described in details in following section 3.1.2.

One may express favour or dislike towards a target by using positive or negative language. We rely on sentiment analysis in determining the stance of an author towards a target. Making an assumption that the stance towards a target might change across different topics, we model stance within topics (subreddit categories). Thereby, we hope to catch emotional patterns typical for authors, patterns of sentiment they express while mentioning certain named entities.

For example, given following target (named entity) in Table 3.1 - Hillary Clinton and a sentiment score of context it is mentioned in, we might suppose

category	target	post text	context sentiment
News and Politics –> Republican	Hillary Clinton	...Besides, Hillary Clinton also has some weaknesses. ...	-0.3612

Table 3.1: Example of entity context sentiment in a CMV post.

that the author's stance towards Hillary Clinton in a current discussion in topic 'Republican' is rather negative.

After extraction of named entities along with context sentiment scores from all author's posts belonging to a certain category, the median value of all scores towards the same named entity is selected as a score of author's stance. Thus, each author is represented as a set of entity vectors with median stance score for every topic she ever has written a post to. We hypothesise that these feature vectors show a tendency that the stance of the user follows. An example of the distribution of median sentiment scores for two different authors in the same subreddit 'personalfinance' is shown on the Figure 3.1.

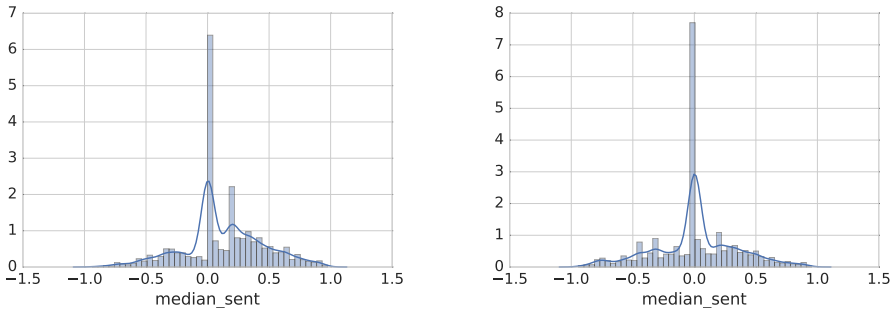


Figure 3.1: Example of entity sentiment scores distribution of two CMV users in the same 'personalfinance' subreddit.

Extracting Entities

For purposes of extraction of entities was selected the Link Probability approach [Milne and Witten, 2008] exploiting Wikipedia statistics and mappings between Wikipedia articles. The Wikipedia's structure provides a set of truly useful features for generating candidate entities, such as article pages (entities), redirect and disambiguation pages, and "wikilinks" interconnecting Wikipedia articles.

- *Wikilinks* are the blue and red hyperlinks between Wikipedia articles. External hyperlinks are not wikilinks, they are not considered in the

entity linking process. Each link has a *target* – page title and an *anchor* text – text of the link itself. In many cases, but not always, target and anchor match. By collecting all variations of anchor text for an entity, it is possible to detect common ways of referring certain entities such as "Big Apple" for New York City.

- *Entities* are wikilink targets – article pages, the encyclopedic content. Entities are represented by the titles of the articles.

We employ the baseline retrieval model of Odijk et al. [2013]. *Prior probability* defines likeliness of the anchor text a to link to Wikipedia article w, also known as *commonness* as introduced by Medelyan et al. [2008] "the extent to which they are well-known".

$$P_{prior}(w|a) = \frac{|lnk_{a,w}|}{|lnk_a|}$$

where w is a Wikipedia article, a anchor text, $lnk_{a,w}$ set of links with anchor a and target w, and lnk_a all links with anchor text a.

Keyphraseness "is defined as the number of Wikipedia articles that use it as an anchor, divided by the number of articles that mention it at all." [Milne and Witten, 2008] An "estimate of the probability that a phrase is selected as a keyphrase for a document" [Mihalcea and Csomai, 2007] or "the probability of being a candidate" [Medelyan et al., 2008].

$$P_{keyphrase}(a) = \frac{DF(lnk_a)}{DF(a)}$$

where a is the anchor text, $DF(.)$ document frequency, $DF(lnk_a)$ is the number of Wikipedia articles where a is used as the anchor text of a link and $DF(a)$ the number of Wikipedia articles containing the text a at all.

As a basis, the initial code of an open-source tool Semanticizest[3] was used for entity linking tasks. We adopted it for our needs and added support of integration with Wikidata to provide the mapping from Wikipedia articles and Wikidata entities, to make use of some additional entity properties provided by Wikidata. As input for entity linking model the Wikipedia dump from March 2018 (containing 5,773,600 articles) and Wikidata dump version from January 2018 containing 42,306,023 entity pages.

[3] `https://github.com/semanticize/semanticizest`

To be able to compute proposed author entity features for such a large dataset as Reddit we employ cluster computing framework Apache Spark[4] running in Webis Research Group Betaweb cluster[5]. Implementation of the feature extraction logic consists of 5 separate Spark jobs organised for convenience into a pipeline. The main steps are taken to process each post to extract named entities:

1. tokenize post content into sentences;
2. tokenize words in a sentence;
3. POS-tag and lemmatize the words;
4. remove stop words and punctuation;
5. run entity linking model on each sentence;
6. compute sentiment score for each sentence.

In the end, for each author, we have a list of topics (subreddit categories) with target entities along with median context sentiment where they were mentioned by the author.

Explorative Web-interface Tool

For convenience, the ad hoc web-based user interface (Figure 3.2) was developed to support exploration and analysis of CMV dataset annotated with named entities. With the help of the tool we adjusted the threshold of *Commoness* score and appended a number of additional steps to reduce noise, for example, filter out entities with stop words as titles and common words representing entities (such as the musician called "Like").

Additionally, for purposes of author similarity analysis, an index for searching most similar authors to a given one was constructed and integrated into the web-based UI tool. We employed PySparnn [6] tool which implements approximate nearest neighbours search for sparse data.

[4] https://spark.apache.org/

[5] `https://webis.de/facilities.html`

[6] `https://github.com/facebookresearch/pysparnn`

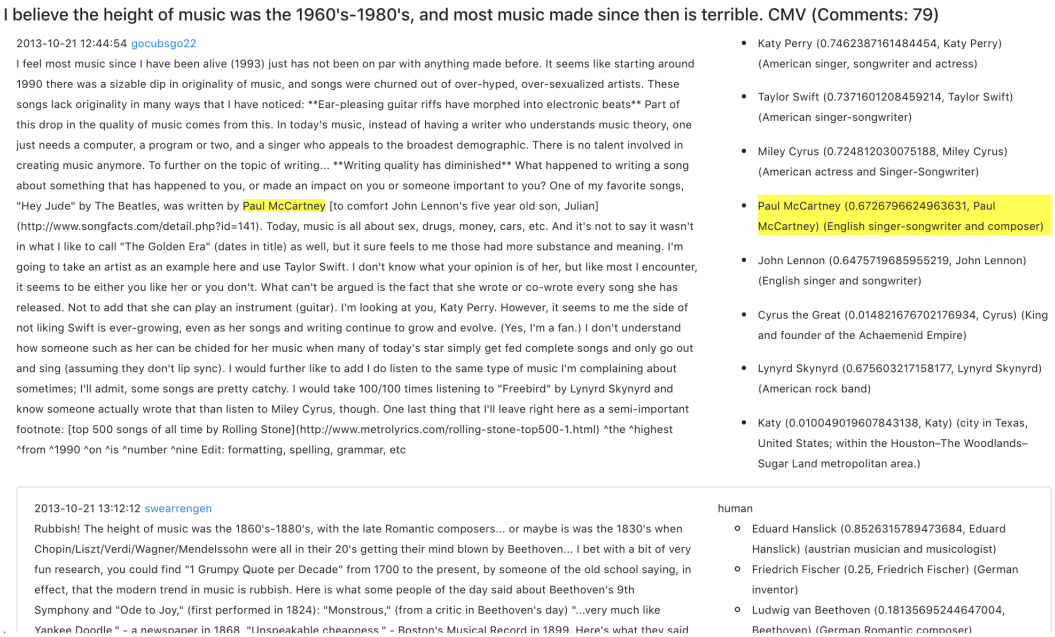


Figure 3.2: Web UI tool for exploration of annotated CMV dataset and similar authors search.

3.1.3 Personality Traits

For the third class of author features, we employ Linguistic Inquiry and Word Count (LIWC) text analysis software tool (2.3) API. It is widely used [Pennebaker et al., 2003], [Tumasjan et al., 2010], [Schwartz et al., 2013] and approved by years of extensive research Pennebaker et al. [2001], Tausczik and Pennebaker [2010b] to assess emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary. The LIWC master dictionary contains almost 6,400 words, word stems, and selected emoticons. It outputs the percentage of words falling into over 80 linguistic, psychological, and topical categories. The web-services of LIWC API provides additional 59 feature categories based on initial word-counts representing personality traits including Big Five, 16PF.

As it is recommended to use as input the texts at least 300 words of length[7], we extracted from ChangeMyView data a set of 9709 authors with first 1000 words of text content of their posts. The authors with a total length of posts less than 300 were excluded. The selected posts were sorted by creation date, merged and then the first chunk of required length were used as input

[7]The upper boundary for used academic licensed version of the API set to 1000 words per request (document) and 10000 total requests

for the API. We make an assumption that earlier posts tend to be less biased in terms of the language used as overtime authors get experience and may use steady patterns in argumentation resulting in diminution of emotional indicators.

Taking raw text as an input the API returns following features:

```
['receptiviti_scores': ['warnings', 'percentiles', 'raw_scores'],
'personality_snapshot', 'content_date', 'content_tags',
'language', 'communication_recommendation', 'user',
'emotional_analysis': ['facets', 'emotional_tone'],
'content_source', '_links', 'liwc_scores', 'organa_scores', 'author']
```

The next three properties are used to construct the personality traits feature-vectors:

1. `liwc_scores` – LIWC categories;
2. `receptiviti_scores.percentiles`: weighted personality traits scores;
3. `receptiviti_scores.raw_scores`: raw personality traits scores.

Five categories from `'receptiviti_scores.percentiles'` corresponding to Big Five (2.4) traits are excluded into separate feature vector.

In the end, for each author the following features representing personality traits are prepared: bigfive, `liwc_scores`, percentiles and `raw_scores`.

3.1.4 Author Similarity

Having author features precomputed the following step of calculation of author similarities was designed. The algorithm requires an input arbitrary pair of authors from the dataset and returns a final set of similarity features. For each type of feature individual similarity score is returned, thus for fine-grained features, separate scores for all 720 subcategories and 20 top-level categories are provided.

We experimented with several similarity measures such as cosine similarity, Jaccard and Euclidean distance, but found no noticeable improvement over cosine similarity only. Cosine similarity for all author feature-vectors is computed as L2-normalized dot product of the vectors A, B as shown in Equation 3.1:

$$sim(A, B) = \frac{\Sigma_{i=1}^{n} A_i * B_i}{\sqrt{\Sigma_{i=1}^{n} A_i^2} * \sqrt{\Sigma_{i=1}^{n} B_i^2}} \tag{3.1}$$

In the case of fine-grained category features, cosine similarities of feature-vectors of common categories are calculated, author similarities of categories beyond the intersection considered to be equal to zero.

3.1.5 Post-level and Interplay Features

We compute some of the most powerful textual features implemented by Tan et al. [2016]. Following classes of features are introduced: the ones based only on the replies text such as a number of words and the other representing interplay between the replies and the original post. Number of words turns out to be a very powerful feature in spite of simplicity, as longer replies can convey more information and be more explicit as confirmed by O'Keefe [1997].

argument-only features	interplay
# words OP	words common all
# words challenger root	words common content
# words challenger path	words common stop
	words jaccard all
	words jaccard content
	words jaccard stop
	words op all
	words op content
	words op stop
	words reply all
	words reply content
	words reply stop

Table 3.2: Textual features of pairs dataset.

For argument-only features, two scenarios are recognized resulting in two feature types. The first case when only the opening argument is considered – "root-reply" and other when the later child comments may contain more important arguments – "path", that contains all comments by root-challenger in a subtree. Thereby, three-word number features are computed for each sample pair as shown in Table 3.2.

The interplay between arguments and original posts are captured by similarity metrics based on word overlap as implemented by Tan et al. [2016]. We computed the proposed four features based on the number of unique words in common between the argument (A) and the original post (O):

- number of common words: $|A \cap O|$,
- reply fraction: $|A \cap O|, |A|$,

- OP fraction: $|A \cap O|, |O|$
- Jaccard: $|A \cap O|$.

The same way suggested by Tan et al. [2016], for all four types of intersection features above we used three different word sets: stopwords, content words and all words. Table 3.2 shows all types of textual features computed for samples of pairs.

3.1.6 Summary

In conclusion, all approaches used to design author level features we highlight three main feature classes: *Interests*, *Beliefs* and *Personality traits*. For interests and beliefs features, two levels were introduced: "flat" and "fine-grained" features. Where, in the case with fine-grained, users are represented with separate feature vectors for each of categories. Personality traits features include 85 psychological word-categories and 59 other personality traits including Big Five and 16PF (2.2.1).

Additionally, Principal Component Analysis (PCA)[8] was applied to several feature groups to reduce the amount of category features. Within each top- and sub-category PCA vectors of length 5 were computed using the incremental approach provided by Sklearn[9] machine learning library. The fine-grained groups of features were used as input for PCA algorithm: subreddit-features within (top-)categories, entity/sentiment features within (top-)categories.

We also experimented with Word embeddings[10] using as input text of author posts, category titles and raw lists of entities, but found no improvement over other proposed author features.

As shown in Table 3.3, altogether there are 39 groups of author features arranged by types. We evaluate the performance of all feature groups in two experiments described in detail in section 4.1 and section 4.2).

[8] `https://en.wikipedia.org/wiki/Principal_component_analysis`
[9] `https://scikit-learn.org/stable/`
[10] `https://en.wikipedia.org/wiki/Word_embedding`

	feature	group label	type
1	words		
1.1	# words root (all/content/stop words)	words	post
1.2	# words path (all/content/stop words)	words	
2	interplay root		
2.1	# common words (all/content/stop words)	interplay_root	
2.2	reply fraction (all/content/stop words)	interplay_root	
2.3	OP fraction (all/content/stop words)	interplay_root	
2.4	Jaccard (all/content/stop words)	interplay_root	
3	interplay path		
3.1	# common words (all/content/stop words)	interplay_path	
3.2	reply fraction (all/content/stop words)	interplay_path	
3.3	OP fraction (all/content/stop words)	interplay_path	
3.4	Jaccard (all/content/stop words)	interplay_path	
4	beliefs similarities		author
4.1	subreddit (raw, tfidf)	belief_ssim	
4.2	category (raw, tfidf)	belief_csim	
4.3	entity (raw, tfidf)	belief_esim	
5	subreddit similarities within (top-)categories		
5.1	subreddit categories (tfidf)	belief_sim_scatt	
5.2	subreddit categories (raw)	belief_sim_scatr	
5.3	subreddit categories pca5 tfidf	belief_sim_scattp	
5.4	subreddit categories pca5 raw	belief_sim_spcatrp	
5.5	subreddit topcategories pca5 tfidf	belief_sim_stcattp	
5.6	subreddit topcategories pca5 raw	belief_sim_stcatrp	
6	entity similarities within (top-)categories		
6.1	entity categories tfidf	belief_sim_ecatt	
6.2	entity categories raw	belief_sim_ecatr	
6.3	entity categories pca5	belief_sim_ecatp	
6.4	entity topcategories tfidf	belief_sim_etcatt	
6.5	entity topcategories raw	belief_sim_etcatr	
6.6	entity topcategories pca5	belief_sim_etcatp	
7	entity/sentiment similarities within (top-)categories		
7.1	entity/sentiment categories	belief_sim_secat	
7.2	entity/sentiment topcategories	belief_sim_setcatp	
7.3	entity/sentiment categories pca5	belief_sim_setcat	
7.4	entity/sentiment topcategories pca5	belief_sim_secatp	
8	traits similarities		
8.1	bigfive	traits_sim	
8.2	percentiles	traits_sim	
8.3	raw_scores	traits_sim	
8.4	liwc_scores	traits_sim	
9	traits features raw		
9.1	bigfive	traits_feat_b	
9.2	percentiles	traits_feat_p	
9.3	raw_scores	traits_feat_r	
10	traits features raw difference		
10.1	bigfive	traits_feat_bdif	
10.2	percentiles	traits_feat_pdif	
10.3	raw_scores	traits_feat_rdif	

Table 3.3: Summary of all author-level and textual features groups.

3.2 Pairs Dataset

To discuss the structure of CMV debates and experimental setup we use the following terminology:

- *submission* – original post of discussion stating the problem; root node for all posts in a discussion;
- *delta-submission* – a submission with at least one delta-awarded comment;
- *comment, node* – a submission or comment, node of a CMV discussion tree;
- *positive, negative comment* – for convenience, we consider comments awarded Δ as positive and not-awarded Delta as negative, respectively;
- *author/user* – any user of CMV including OP;
- *challenger* – any member of discussion besides the OP;
- *OP* – the original poster, author of discussion;
- *Discussion tree* or *submission* is the original post containing statement of the problem and collection of all the comments of a separate discussion;
- *Root-reply, top-level comment* – a first level comment, direct reply to the original post. Authors of root-replies are root-challengers;
- *Discussion subtree* – subset of comments of a discussion, a root-reply and all its child comments;
- *Terminal comment* –a leaf comment in a discussion branch;
- Discussion *path* or *branch* contains all comments starting from root-reply to a terminal comment and contains all comments by root-challenger.

Several preliminary steps were done to clean the raw corpus before continuing with feature construction. For purposes of optimization and since for both of the tasks we analyse only authors of CMV subreddit, initial Reddit dataset was filtered leaving only users that have at least one submission or comment in ChangeMyView community.

In addition, a standard footnote from CMV moderators was removed and all submissions and comments in the Reddit dataset meeting following conditions were filtered out:

1. the record is not a valid JSON;
2. properties 'author' and 'content' not exist in the record;
3. the post content is empty or equals '[removed]', '[deleted]';
4. author in botlist or equals '[deleted]';
5. duplicates of comments with earlier datetime of 'retrieved on';
6. for comment records, no valid link to parent record.

The employed Reddit corpus is stored in JSON lines[11] format, where each submission and comment is a separate JSON object. To extract samples for the paired dataset the original hierarchical structure of submissions has to be recovered. We divided this task into two steps.

First, the original comment trees are reconstructed. For this purpose, we employ `'parent_id'` and `'link_id'` properties of JSON objects from the raw corpus: first connects the node to the parent submission, while the second points to parent node (submission for root-replies and parent comment for child replies down along the discussion subtree).

In total, 65,169 discussion trees were constructed out of raw corpus, some statistics are given in Table 3.4. Full discussion tree for each submission was created, preserving all posts by bots and not applying additional filtering in order to preserve the original hierarchy of discussions.

# discussion trees	# nodes	# OPs	# uniq. authors
65,169	3,449,917	28,722	155,337

Table 3.4: Discussion trees dataset statistics.

Next, paired samples were extracted from the discussion trees. In order to reproduce results for argument-level features and evaluate the performance of author features, we adopted the methodology of Tan et al. [2016] to create the pairs dataset.

We focus on top-level comments since the root reply is what usually initiates a new branch of argumentation and defines if the OP will join the discussion. To avoid insignificant cases, we apply several conditions for discussion candidates:

[11] `https://en.wikipedia.org/wiki/JSON_streaming#Line-delimited_JSON`

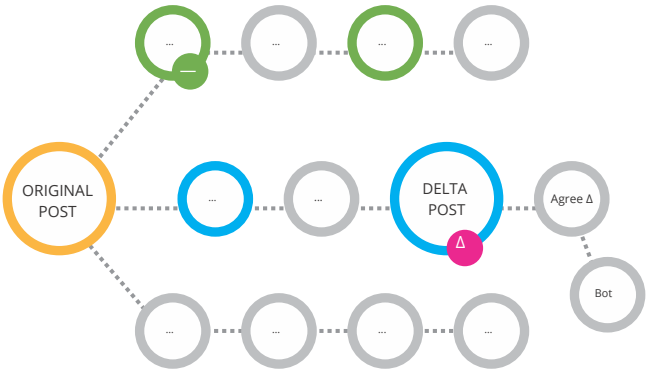


Figure 3.3: The CMV submission. Replies by different authors are colour-coded. The delta-awarded discussion subtree and the selected negative subtree are highlighted.

- only discussions where at least one Δ was given by the OP are taken into account;
- to be sure that enough arguments were given to original post, only discussion trees with more than 10 root-challengers are accepted;
- only comments by root-challengers are considered as candidates;
- root-replies by authors with name '[deleted]' or containing in the botlist are removed;
- to avoid cases where replies are not arguments but clarifying questions, root-replies of length less than 50 words are removed;
- although all challengers are allowed to award Δ to comments they believe are persuasive, we only consider deltas given by the OP so far as we focus on OP's persuasion;
- for accepted subtrees, all replies by a root-challenger in the line of argument are extracted (refers to "path"-case in 3.1.5).

Similarly to Figure 2.2, the Figure 3.3 shows an example of a submission structure with three branches of discussion (subtrees). The coloured replies are the replies that will be selected as samples for the pairs dataset. We extract the subtree with Delta-awarded reply saving all replies by the top-challenger as the whole line of argument. From the other two "unrewarded" subtrees we select the one most similar

to the original post in terms of the language used. The comments from this branch are highlighted with green colour.

The recursive Algorithm 1 was implemented to traverse the discussion trees and extracts replies-candidates. For each accepted as candidate positive path, we search for the negative branch within the same discussion tree which is the most similar.

Algorithm 1 Recursive pairs extraction from subtrees

```
1: procedure GENERATE_PAIRS(comment, submission, op, pairs, context)
2:     author ← comment[author]
3:     direct_op_reply ← is_direct_op_reply(comment)
4:     interplay ← is_interplay(comment, author, op)
5:     delta ← is_delta(comment, op)
6:     if is_terminal(comment) or not is_ops_reply(comment, op) then
7:         return pairs
8:     end if
9:     if direct_op_reply and (not interplay or (interplay and
10: delta)) then
11:        APPEND(context[op], direct_op_reply)
12:        APPEND(pairs, tuple(submission, comment))
13:    end if
14:    APPEND(context[author], direct_op_reply)
15:    children ← comment[children]
16:    for c ∈ children do
17:        GENERATE_PAIRS(children[i], submission, op, pairs, context)
18:    end for
19:    return pairs
20: end procedure
```

The similarity between branches is measured based on Jaccard similarity in root-replies with removed stopwords:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where A and B are sets of words in the two root-replies. In this way, two lexically similar lines of argument are selected for the task of prediction which of them is a successful one.

'op'	'author', 'created_utc', 'id', 'num_comments', 'selftext', 'subreddit', 'title', 'features': ['beliefs', 'interests', 'personality_traits']
'a'	'author', 'comments', 'delta', 'features': ['beliefs', 'interests', 'personality_traits']
'similarity_features'	'author', 'path', 'root'
'delta'	True/False

Table 3.5: The structure of the samples in constructed CMV dataset.

Total of 16,050 pair samples was extracted. Table 3.5 shows an example of the structure of a sample of a pair. For each pair of information about the original post is provided along with author features, all comment by a challenger together with her author-level features and similarity features of the authors in the sample. In comparison with the dataset employed by Tan et al. [2016] in their experiments, the constructed dataset comprises of over than four times more samples respectively.

Having for each pair positive and negative samples were collected, allows us to create a balanced dataset. Out of period from January 2013 till September 2017, we put away approximately last 6 months (discussions with creation date after `2017-04-01`) containing 3,554 for heldout evaluation dataset, that leaves 12,496 samples for the training dataset.

Figure 3.4 shows how often in extracted pairs the same authors are encountered. Although in most cases the same two authors in the same pair occur only one time, the number of authors who interacted two times and more are significant.

Additionally, special "historical" author features were computed to analyse how the amount of information about a user is correlated with models accuracy. For this purpose, the number of words from all author posts across Reddit communities was summed up. This way, for each author in a pair sample a feature representing his or her "experience", was provided. As the next step, for each pair following extra features were computed:

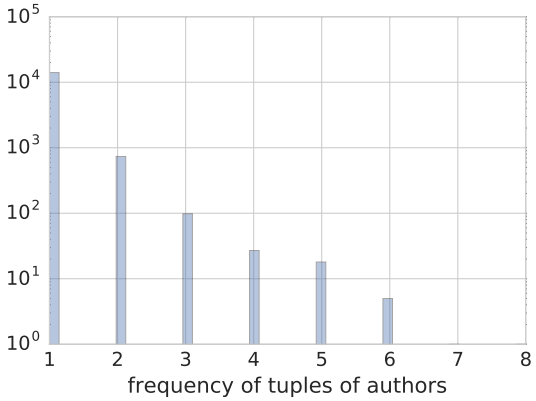


Figure 3.4: Frequency distribution of combinations of authors in the pairs dataset.

- the minimum history score: $words\_min = min(|W_{author_1}|, |W_{author_2}|)$, where W_{author_i} is the number of words in posts by an author across all subreddits;
- similar to minimum, maximum history score: $words\_max = max(|W_{author1}|, |W_{author2}|)$;
- and average of $words\_min$ and $words\_max$ history scores.

Both total post words and the number of stop words were computed. These values were used for stratification of the heldout dataset at the evaluation step. Figure 3.5 shows the distribution of minimum experience score in the dataset.

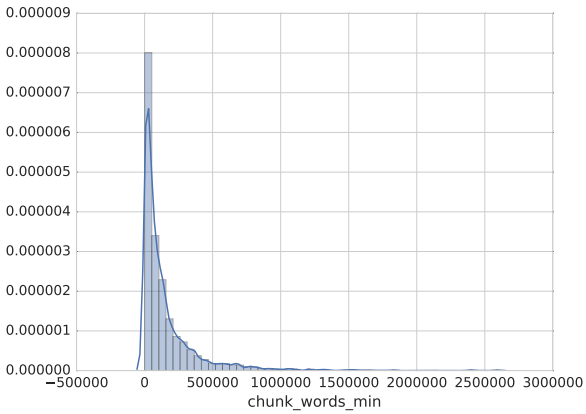


Figure 3.5: Distribution of $words\_min$ historical feature in the pairs dataset.

Subreddits Statistics

While constructing subreddit topics -based features we collected some general statistics on the distribution of the most popular communities topics on Reddit. On Figure 3.6 top 30 author frequencies of subreddits are shown.

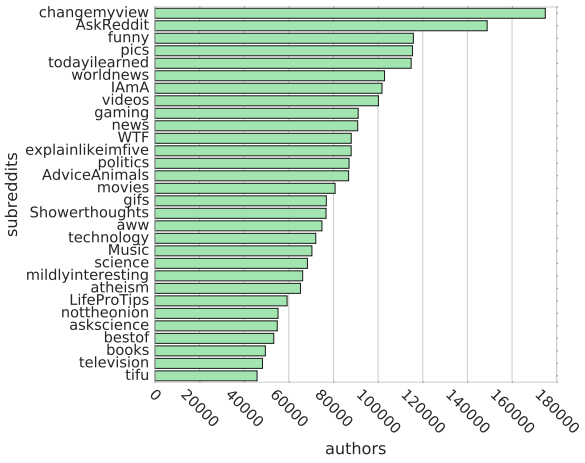


Figure 3.6: Number of unique authors in top 30 subreddits.

Figure 3.6 shows frequency distribution of authors in different subreddits. As results of filtering out all users except CMV authors at the pre-processing step, the highest frequencies have ChangeMyView subreddit and Discussion category to which CMV belongs respectively. Besides that, frequency distributions of authors shown on figures 3.6 3.7 follow the same patterns as in not filtered dataset.

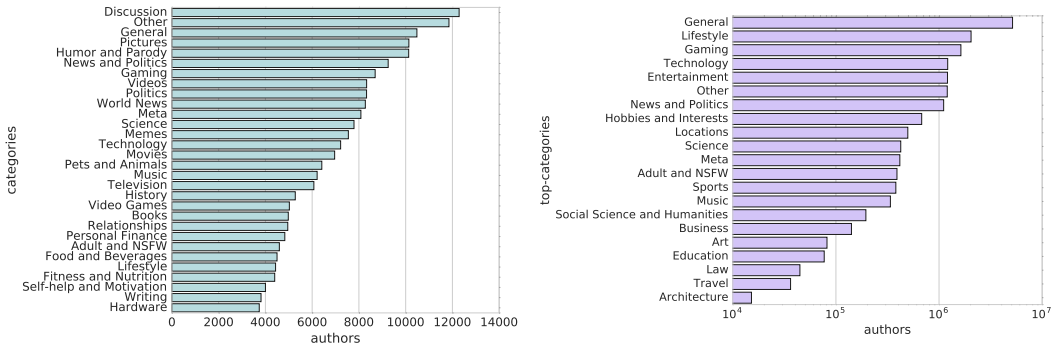


Figure 3.7: Number of unique authors in 30 most frequent subreddit subcategories and top-level categories.

Chapter 4

Evaluation

In this chapter, we describe the settings for two experiments. First, we evaluate author-level features by conducting the task of prediction of the persuasive comment from a pair of replies. In the second task, we classify the original poster's opinion as malleable or resistant using personality traits features and compare the performance with the post-level textual features.

We use Logistic Regression classifier in four different experiment settings: all features together, each group separately, `'post'` group + each of other groups, ablation setting and all combinations of group types. Finally, we evaluate the models built within each setting on the stratified heldout dataset.

4.1 Predicting the Successful Argument

In our first task, we try to predict the persuasive comment which was awarded Delta given a pair of comments to the same original post. Each of the comment pairs is represented by the features discussed in chapter 3 covering the original post and both comments. We explore how the similarity of user interests, beliefs and personality traits is correlated with persuasion success in online debates. The dataset used for the experiment is described in detail in section 3.2. Each pair sample in the dataset was split into negative and positive samples comprising of delta-awarded comment and one not awarded Δ respectively, letting us have a balanced evaluation dataset. In this way, each sample contains two sets of author similarity features and textual features (for the original author and author of each of two replies respectively).

We hypothesise that when given more data about an author the more accurate our predictions should be. In order to confirm the hypothesis, we employ information about the level of author past activity to stratify the testing dataset. The precomputed "historical" author features were used for splitting the heldout dataset into buckets for evaluation. Considering the distribution of values of *words_min* historical author feature we divided testing set into four approximately even buckets. We

test all the models on each bucket separately and report these results in each experiment setting. The Table 4.1 shows the distribution of samples in stratified heldout set.

group	delta	nodelta
<20k	471	430
20k-50k	418	420
50k-130k	447	405
130k+	502	461

Table 4.1: Distribution of positive and negative samples in the heldout dataset split into four buckets by *words_ min* "historical" author feature.

Experimental Setup

The same author and textual features groups were used throughout the experiments as listed in the Table 3.3. Finding a trade-off between the time required for training the models and the optimal amount of feature groups we decided for 9 types of groups. This gives us a total of 511 combinations of group types. The author features are grouped into feature types as shown in Table 4.2.

	type	groups
0	post	words, interplay_path, interplay_root
1	traits_flat	traits_sim
2	traits_feat_b	traits_feat_b, traits_feat_bdif
3	traits_feat_p	traits_feat_p, traits_feat_pdif
4	traits_feat_r	traits_feat_r, traits_feat_rdif
4	belief_sim	belief_ssim, belief_csim, belief_esim
6	belief_ec	belief_sim_ecatp, belief_sim_ecatr, belief_sim_ecatt, belief_sim_etcatp, belief_sim_etcatr, belief_sim_etcatt
7	belief_sc	belief_sim_scatr, belief_sim_scatt, belief_sim_scattp, belief_sim_spcatrp, belief_sim_stcatrp, belief_sim_stcattp
8	belief_sec	belief_sim_secat, belief_sim_secatp, belief_sim_setcat, belief_sim_setcatp

Table 4.2: Feature group types (group labels as in Table 3.3).

We train the models in four different settings: (1) three main feature groups: all, author and post (all textual features), (1) each feature group separately, (2) ablation setting - all feature groups except one, (3) all features together and (4) all possible

combinations of types. The Logistic Regression classifier with L1 regularization was used for training and evaluation and 5-fold cross validation for parameter tuning. We, then, compare the AUC and the accuracy scores on the stratified heldout dataset.

Main Feature Types

In this setting, we trained the models on three main types of features: author-level features, textual features and all author and post feature together. The Table 4.3 shows the performance of each feature levels above for all of four buckets.

The textual features perform significantly better then author-level features in this setting. Even all features groups together are less predictive than post features alone, likely due to high noise in case of using all of 7,258 features. Although, it is an interesting observation that performance on the smallest bucket of the heldout dataset, that implies we have less information about these users, is almost at the same level as for the biggest bucket.

group	strata	accuracy	f1	roc_auc	precision	recall
author	<20k	53.94	58.29	0.56	0.54	0.54
author	20k-50k	**55.13**	56.78	**0.56**	0.55	0.55
author	50k-130k	54.10	55.82	0.55	0.54	0.54
author	130k+	54.72	57.25	0.56	0.55	0.55
all	<20k	58.388	61.85	0.61	0.58	0.58
all	20k-50k	58.23	59.30	**0.61**	0.58	0.58
all	50k-130k	57.51	58.77	0.60	0.57	0.58
all	130k+	**58.57**	60.84	0.61	0.58	0.58
post	<20k	58.16	63.57	0.61	0.58	0.58
post	20k-50k	56.80	61.40	0.62	0.57	0.57
post	50k-130k	57.28	62.24	0.61	0.57	0.57
post	130k+	**58.46**	64.15	**0.62**	0.58	0.58

Table 4.3: Performance of all features grouped by the feature type on stratified heldout dataset.

Feature Groups

Next, we evaluate each of the 30 feature groups alone. We report the accuracy on each bucket of heldout dataset (Figure 4.1).

The textual `interplay_path` feature group performs the best with AUC score of 0.61. At the same time, features representing author personality traits such as `traits_feat_p`, follow close behind with AUC=0.59.

Although only some of the belief feature groups perform better then baseline `words` feature (AUC=0.51), in general, they perform noticeably better on buckets with higher values of "historical" author feature, achieving 0.56 AUC with feature

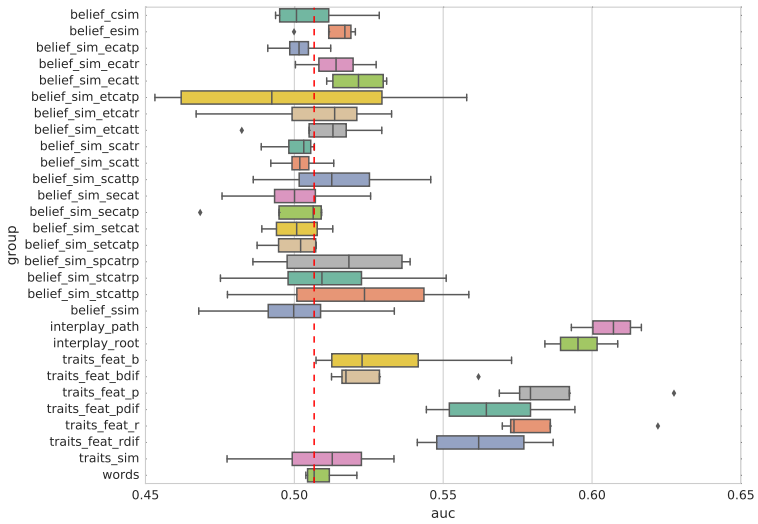


Figure 4.1: Accuracy of feature groups on stratified heldout dataset. Median AUC score of `'words'` feature group used as a baseline.

`'belief_sim_stcattp'`. That is in favour of the assumption that the more data we have about the user, the more accurate our predictions are. However, in total, they still perform worse then interplay features. It seems that beliefs features alone are not very informative and author similarity, when used solely, is not meaningfully correlated with persuasion success. At the same time, traits feature performance, whereas also highly fluctuating from bucket to bucket, beat interplay features achieving AUC 0.63 over 0.61 on the top bucket '130k+' with `'traits_feat_p'` feature.

Post Features Combined with Other Feature Groups

Since `'post'` feature group is indeed a very powerful indicator, we designed a setting to explore how author-level features when combined with textual features, improve the accuracy of classification persuasive comments. For each feature group from listed in Table 4.2 we trained a classifier using features from the group and all textual features of the `'post'` type.

As can be seen on Figure 4.2 in this setting all personality traits and many of beliefs features outperform textual features even though, most pronounced when evaluated on top buckets of the heldout dataset. `'traits_feat_p'` achieves AUC score of 0.66, most predictive among belief features `'belief_sim_stcattp'` 0.64, while all textual features together achieve maximum of AUC 0.62.

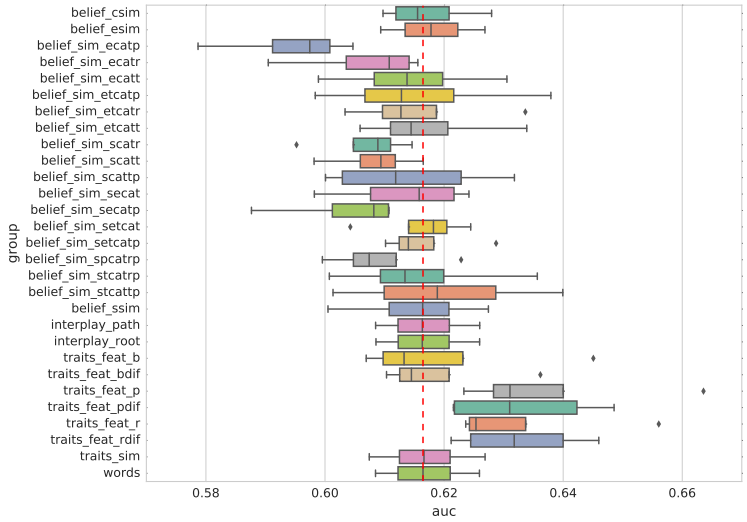
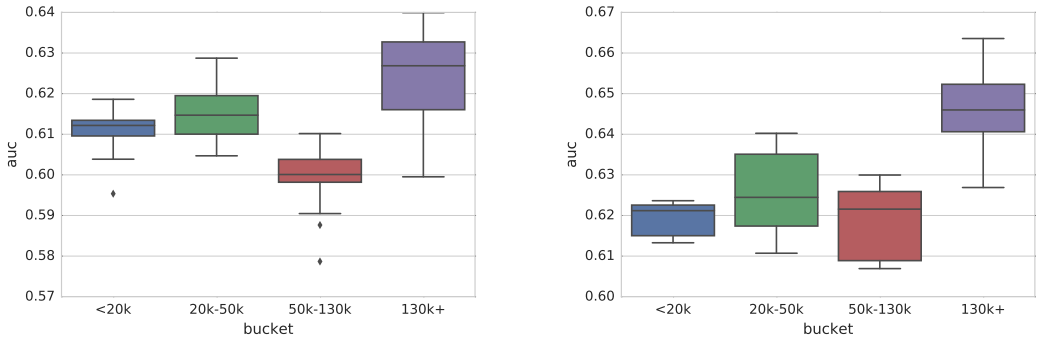


Figure 4.2: Performance of feature groups together with post features.

We provide separate Figure 4.3 for beliefs and traits features showing the performance of each feature group for each of the buckets of the heldout dataset. The graphs reflect the tendency of accuracy improvement with the increase of the data we know about the users. An interesting observation is, that the '50k-130k' bucket drops out the common patterns. We did not have time to investigate it profoundly, but probably there is a certain factor influencing on the authors' data within this range of historical feature values.



(a) beliefs **(b)** traits

Figure 4.3: Performance of all beliefs and traits feature groups combined with post features.

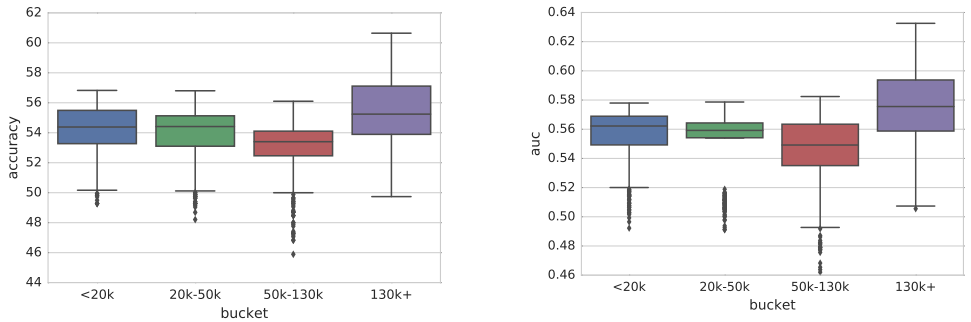
Combinations of Feature Types

Lastly, we prepare an experimental setting letting us stepwise evaluate all possible combinations of 9 feature types. For each combination of feature types, the Logistic Regression classifier was trained on the subset of all features belonging to each type as listed in Table 4.2. For this setting, we omitted parameter tuning with cross-validation in order to decrease the time required and used the default parameters.

group	strata	accuracy	f1	roc_auc	precision	recall
traits_feat_p +post +traits_flat +belief_sim	130k+	**62.20**	**64.67**	**0.67**	**0.62**	**0.62**
traits_feat_p +post	130k+	61.27	64.31	0.66	0.61	0.61
post +traits_flat +belief_sim	130k+	59.19	63.38	0.63	0.59	0.59
post +belief_sim	130k+	58.99	62.99	0.63	0.59	0.59
traits_feat_p	130k+	60.23	64.17	0.63	0.60	0.60
post	130k+	58.46	64.16	0.63	0.59	0.58

Table 4.4: Performance of top representative feature combinations on '130k+' bucket of heldout dataset.

The top performing combination `'traits_feat_p +post +traits_flat +belief_sim'` achieves AUC 0.64 on the whole heldout dataset and 0.67 on '130k+' bucket respectively. The combination includes powerful textual features along with beliefs feaures and traits: `'traits_feat_p'` group contains 59 different personality traits features, `'traits_sim'` – four similarity scores for each of traits feature vectors and `'belief_sim'` holds three similarity scores for "flat" author beliefs feature vectors. In Table 4.4 we provide results for several most important feature combinations to highlight how the performance of the classifier increases with addition of author-level features.



(a) accuracy

(b) AUC score

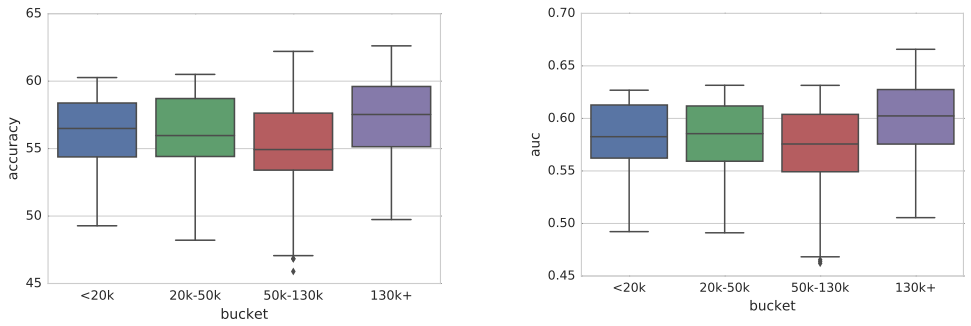
Figure 4.4: Performance of all author feature combinations excluding with textual features.

Additionally, we compare the performance of all combinations of author features

without textual on each heldout bucket as shown in Figure 4.4 and author feature combinations together with textual features (Figure 4.5).

We can observe the same pattern as in section 4.1 that is in favour of the assumption about the more data we have the more accurate are our predictions. Although, it is not that strong in the case when we consider all features together.

When combined with post features combinations of author features perform noticeably better and with less noise (outliers). At the same time, the observable improvement in the case when they are combined, let us assume that when the context of argumentation is known, the predictive power of author-level features significantly improves the accuracy of the classifier. We speculate that when two challengers in a discussion provide similar argumentation, the original poster will rather be in favour of the author more similar to herself in terms of beliefs and personality traits. Anyway, without the argumentation context, the author features cannot provide stable performance.



(a) accuracy

(b) AUC score

Figure 4.5: Performance of all author feature combinations together textual features.

Conclusion

We evaluate author-level features by conducting the task of the prediction of the persuasive comment from a pair of replies. The four different experiment settings are used: all features together, each group separately, `'post'` group + each of other groups, ablation setting and all combinations of group types. The created models are evaluated within each setting on the stratified heldout dataset. Our results show a noticeable improvement in the case when textual features are used together with author-level features. Our hypothesis about that given more data about the author the predictions should be more accurate was not confirmed in complete, although

showing the tendency of improvement of prediction with the increase of the amount of input data.

4.2 Opinion Malleability

In the second task, we, given the original post, try to predict if the opinion expressed by the author is resistant or malleable. Each original post is represented by the personality traits and textual features discussed in chapter 3. We explore how personality traits are correlated with the malleability of the opinion.

For this experiment, the internal structure of a discussion is not taken into account and only the discussions with at least one reply from the OP are considered in order to ensure that he or she has engaged into discussion.

As long as CMV submissions alone are taken into account, author similarity features are not meaningful in this setting as a single author is present in samples – the original poster of a submission. Hence, only the personality traits and textual features are evaluated.

For this task, we report results for two datasets, extracted from our version of Reddit corpora (CMV) and the other provided with Tan et al. [2016] (WinArgs dataset). It is not possible to compute author features for all users in WinArgs dataset since it was crawled over three years ago and some of the users' accounts on Reddit were deleted during this time leaving unrecognizable `'[deleted]'` user names. This lead to a reduction of dataset size to 3,934 samples in the training set and 780 in the heldout.

Preparing Dataset

For this experiment different dataset was extracted from the ChangeMyView corpora. Using the discussion trees dataset described in section 3.2 as an input data, we only leave submissions meeting the following requirements:

- only discussions where the OP replied at least once are taken into account;
- to be sure that enough arguments were given to original post, only discussion trees with more than 10 root-challengers are accepted;
- the content of the record is not empty or equals `'[removed]'`;
- to rule out poorly expressed posts, submission of length less than 50 words are removed;
- the author property does not equals `'[deleted]'`.

dataset	# submissions	# delta	# uniq. authors
CMV	8,121	3,186	4,641
WinArgs	4,714	2,244	2,348

Table 4.5: Submissions dataset statistics.

Some general statistics of both datasets are shown in Table 4.5.

The Figure 4.6 displays the distribution of the number of Delta-awarded replies per discussion in the dataset. As we can see, although many of the original posters usually give only one Δ, almost in half of the cases, multiple challengers manage to change the OP's opinion. It would be interesting to explore if there is a predisposition for original posters with certain personality traits to give more Deltas.

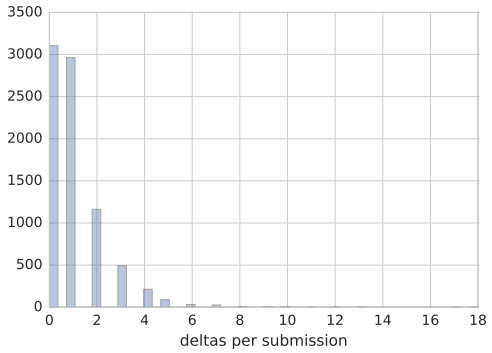


Figure 4.6: Distribution of number of comments awarded Δ per submission.

Features

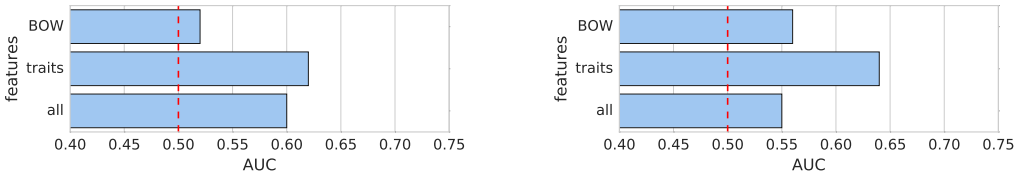
Unlike the first task, in this case, it is not possible to compute such features as an interplay or author similarities. Thereby, Bag-Of-Words (BOW) feature was computed as representative of the class of textual features and personality traits features from author features (3.1.3).

For BOW features minimum frequency parameter was similarly to set up of Tan et al. [2016] chosen equal 5 and all feature vectors were standardized by removing the mean and scaling to unit variance. Following traits features were used (as in Table 3.3): `'traits_feat_b'`, `'traits_feat_p'`, `'traits_feat_r'`.

Results

We use Logistic Regression classifier in three experiment settings: all features together, BOW features and personality traits features. The Logistic Regression classifier with L1 regularization was used for training and evaluation. We choose a 5-fold cross validation for parameter tuning the same way as in section 4.1.

BOW features perform just over the random baseline. As Tan et al. [2016] mention, it is indeed a very difficult task to predict if the original poster will change her mind. They report results of the pilot study where human annotators perform no better than at a chance level (50%). Experimenting with both datasets we obtain the same level of performance with BOW features as Tan et al. [2016] report.



(a) Pairs dataset **(b)** WinArgs dataset

Figure 4.7: Opinion malleability prediction performance on both datasets.

As shown on Figure 4.7, on both datasets personality traits features outrank textual features by almost 10 percent. We observe significant improvement of accuracy of prediction malleability of the OP's opinion, AUC 0.64 over 0.53. Combined together, all features significantly increase noise, especially in the case with smaller WinArgs dataset where the performance of all features combined is lower than BOW features alone.

We perform analysis of feature importance and correlation between personality traits features and malleability of the point of view of the original poster (OP). Pearson coefficient correlation matrix shown on Figure 4.8 gives some insights in how Big Five personality traits are correlated with the probability of being persuaded.

We observe several interesting patterns in the correlation between personality traits features. It expresses the emotional aspects of being persuaded: neurotic people are more likely to experience negative feelings such as anxiety, worry, fear or anger and probably will question the persuasion attempts and be more hesitant in changing the opinion.

At the same time, agreeable people, as defined by the personality trait, tend to be more sympathetic, cooperative, and considerate that may be letting themselves be persuaded to escape the conflict. The negative correlation of opinion malleability and extraversion appears to follow the findings [Akert and Panter, 1988] that it is

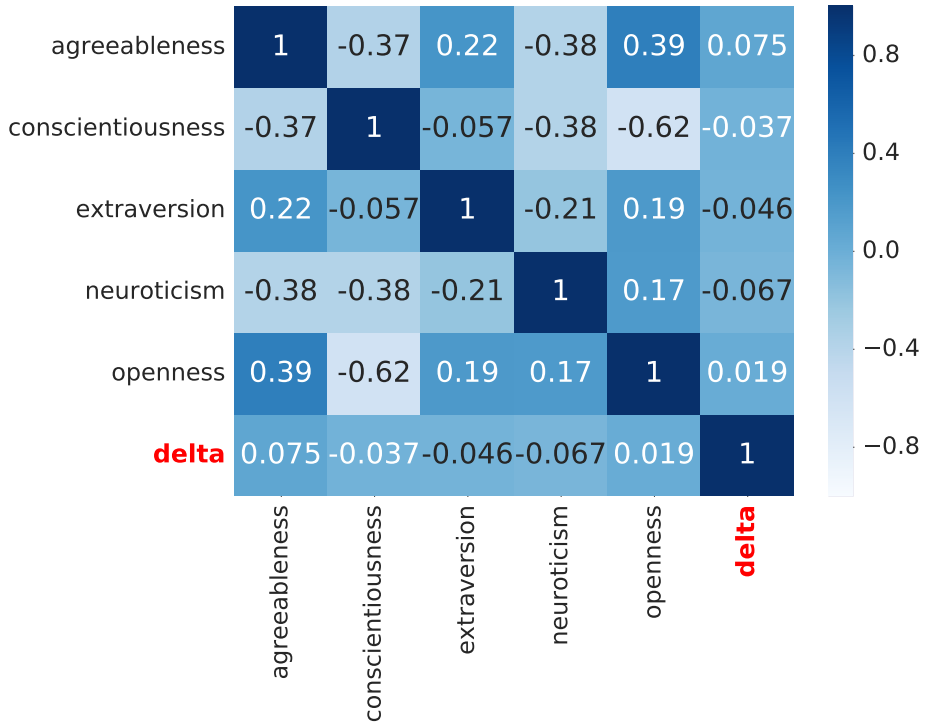


Figure 4.8: Correlation matrix for the OP's Big Five personality traits features and malleability of her opinion in CMV dataset. Here 'delta' means that at least one argument in the discussion was awarded Δ by the OP.

highly correlated with confidence. Thereby, more extraverted people may tend to be surer of themselves, and harder to be persuaded.

The openness, as part of the factual component of the process of being persuaded shows that people who tend have an active imagination, attentiveness and intellectual curiosity and capable of assuming different points of view, and are thus more likely to be persuaded successfully. Conscientiousness reflects traits of being careful, diligent and implies having a serious and responsible attitude towards tasks, that might explain a lower likelihood of persuasion success.

In general, as one could assume intuitively, the more the person is open to the new and sympathetic, the higher chance she would listen more carefully to other's arguments and more likely would be persuaded. Our finding from correlation personality traits with the opinion malleability confirm such hypothesis: the two most positively correlated traits are *agreeableness* and *openness*, while *consciousness*, *neuroticism* are negatively correlated with opinion malleability.

4.3 Summary

In this chapter, we describe two experiments conducted for evaluation of proposed author-level features. The first section describes the experimental settings designed for the task of prediction the persuasive reply out of a pair of comments from the same discussion. We perform various feature performance analysis and report results on stratified heldout dataset.

The succeeding section gives insights into the task of classification of the original opinion of a discussion as resistant or malleable. For this experiment performance of personality traits features are compared with textual features of the posts themselves. In addition, correlation analysis of personality individual personality traits is performed.

Chapter 5

Conclusion and Future Work

In this work, we propose several types of author-level features representing a person's beliefs and personality traits and explore the influence these features have on mechanisms behind the persuasion. In addition, we compute textual features of arguments and compare performance with the author-level features.

In this chapter, we give an overview of the main contributions and present the potential areas of improvement on the current work.

5.1 Contributions

We started our research by formulating our research questions as following: (1) How to model the similarity of people's interests, beliefs, and personality traits? (2) How to operationalize the developed model in terms of identifying the model's attributes automatically? (3) How to demonstrate the impact of the similarity of people's interests, beliefs, and personality traits on their persuasiveness, if any?

To answer these questions, in chapter 3 we analyze user data in the Reddit communities and use it to create new author-level features which represent the person's prior-beliefs and personality traits. We conduct two experiments in order to evaluate the author-level features.

For these purposes, we create two datasets: (1) containing samples of pairs of an original post and two replies with similar argumentation and (2) second comprising of examples of original posts only. We employ the corpora of the Reddit online forum website and its community ChangeMyView which is designed to provide special discussion platform with rules that help to explicitly highlight argumentative replies.

In the first experiment, we predict the most argumentative reply out of a pair of comments from the same discussion. The second experiment designed to classify the opinion expressed by the original poster as resistant or malleable. Our results show a noticeable improvement in the case when textual features are used together with

author-level features. For the task of classification the original opinion, we report significant improvement of the prediction accuracy over state-of-the-art.

To provide answers to our research questions, in chapter 4 we analyze the results of both of the experiments and provide insight into the influence of author features on persuasion success. We explore how personality traits are correlated with opinion malleability and describe our observations.

5.2 Future Work

Our study of the effect of prior beliefs and personality traits on the persuasiveness can be used in different areas such as online advertising and marketing. We assume, through modelling users' beliefs and traits, the level of engagement of potential customers can be increased.

There are many potential ways to improve the current work that is worth to pursue. Most of the methods which are employed in the thesis have the potential for future improvements. Employing state-of-the-art entity linking approaches can provide more robust output for modelling beliefs. Similarly, another direction of improvement is to employ state-of-the-art stance classification.

The use of the topic modelling can potentially increase the performance of prediction. We assume, controlling for a topic of a debate can improve the accuracy of prediction of author-level features.

Besides, further evaluation of author-level features can be done using different datasets such as Debate.org[1] portal. Unlikely the Reddit's CMV, the Debate.org portal provides additional information from user profiles such as the list of stances towards major debate topics (for example, Politics: Conservative vs. Democrat). It would be worth to compare the performance of user profile features with automatically constructed out of posts history author-level features.

Finally, the datasets constructed within the work can be used to design analogous studies in order to evaluate human performance in corresponding tasks.

In this work, we confirmed the hypothesis of the presence of the influence of author-level features on mechanisms behind persuasion and proposed a way to employ a person's characteristics in the argumentation analysis.

[1] `https://www.debate.org/`

Appendix A

Appendix 1

Table A.1: Results of the experiment of predicting the winning argument (4.1) in the experimental setting 'each feature group alone' (sorted by AUC score) on the whole dataset.

group	accuracy	f1	roc_auc	precision	recall
interplay_path	**57.82**	**63.85**	**0.61**	**0.58**	**0.57**
interplay_root	57.46	63.97	0.60	0.58	0.57
traits_feat_p	56.39	60.22	0.59	0.56	0.56
traits_feat_r	56.25	59.49	0.59	0.56	0.56
traits_feat_pdif	55.37	59.38	0.57	0.55	0.55
traits_feat_rdif	54.45	58.52	0.56	0.54	0.54
traits_feat_b	53.07	59.18	0.53	0.53	0.53
traits_feat_bdif	52.34	59.32	0.53	0.52	0.52
belief_sim_ecatt	51.66	64.62	0.52	0.51	0.50
belief_sim_stcattp	51.97	61.63	0.52	0.51	0.51
belief_sim_scattp	51.91	61.15	0.52	0.51	0.51
belief_sim_ecatr	51.49	63.05	0.52	0.51	0.50
belief_sim_spcatrp	52.14	60.71	0.51	0.52	0.51
belief_esim	51.86	65.13	0.51	0.51	0.51
belief_sim_stcatrp	50.76	59.71	0.51	0.50	0.50
traits_sim	51.63	66.69	0.51	0.51	0.50
words	51.86	64.73	0.51	0.51	0.51
belief_sim_etcatt	51.46	64.38	0.51	0.51	0.50
belief_csim	51.63	66.99	0.51	0.50	0.50
belief_sim_secat	51.24	62.99	0.50	0.50	0.50
belief_sim_etcatr	51.66	62.27	0.50	0.51	0.51
belief_sim_etcatp	51.49	61.98	0.50	0.51	0.51
belief_ssim	50.45	58.00	0.50	0.50	0.50
belief_sim_ecatp	50.11	57.59	0.50	0.50	0.50
belief_sim_scatt	49.47	60.30	0.50	0.48	0.49
belief_sim_setcat	51.66	62.27	0.50	0.51	0.51
belief_sim_setcatp	51.01	59.45	0.50	0.50	0.50
belief_sim_secatp	50.56	58.02	0.50	0.50	0.50
belief_sim_scatr	50.06	60.39	0.50	0.49	0.49

Table A.2: Results of the experiment of predicting the winning argument (4.1) in the experimental setting 'combinations of group types' (sorted by AUC score) on the whole dataset.

group	accuracy	f1	roc_auc	precision	recall
traits_feat_p +post +traits_feat_r +traits_feat_b	**60.83**	**63.46**	**0.64**	**0.61**	**0.61**
traits_feat_p +post +traits_feat_r	60.78	63.41	0.64	0.61	0.61
traits_feat_p +post +traits_flat +traits_feat_r +traits_feat_b	60.47	63.21	0.64	0.6	0.6
traits_feat_p +post +traits_flat +traits_feat_r	60.47	63.19	0.64	0.6	0.6
traits_feat_p +post +belief_sim +traits_feat_r +traits_feat_b	60.92	63.5	0.64	0.61	0.61
traits_feat_p +post +traits_flat +belief_sim +traits_feat_r	61	63.53	0.64	0.61	0.61
traits_feat_p +post +belief_sim +traits_feat_r	60.97	63.53	0.64	0.61	0.61
traits_feat_p +post +traits_flat +belief_sim +traits_feat_r +traits_feat_b	61.03	63.56	0.64	0.61	0.61
traits_feat_p +post +traits_feat_b	60.64	63.52	0.64	0.61	0.6
traits_feat_p +post	60.64	63.54	0.64	0.61	0.6
traits_feat_p +post +traits_flat	60.16	63.16	0.64	0.6	0.6
traits_feat_p +post +traits_flat +traits_feat_b	60.3	63.26	0.64	0.6	0.6
traits_feat_p +post +belief_sim +traits_feat_b	60.41	63.25	0.64	0.6	0.6
traits_feat_p +post +traits_flat +belief_sim	60.61	63.39	0.64	0.61	0.6
traits_feat_p +post +belief_sim	60.47	63.31	0.64	0.6	0.6
traits_feat_p +post +traits_flat +belief_sim +traits_feat_b	60.61	63.43	0.64	0.61	0.6
post +traits_flat +traits_feat_r +traits_feat_b	59.79	62.58	0.63	0.6	0.6
post +traits_flat +traits_feat_r	59.85	62.71	0.63	0.6	0.6
post +traits_feat_r +traits_feat_b	60.24	63.04	0.63	0.6	0.6
post +traits_flat +belief_sim +traits_feat_r +traits_feat_b	60.19	63.03	0.63	0.6	0.6
post +traits_feat_r	60.16	63.01	0.63	0.6	0.6
post +traits_flat +belief_sim +traits_feat_r	59.9	62.78	0.63	0.6	0.6
post +belief_sim +traits_feat_r +traits_feat_b	60.35	63.16	0.63	0.6	0.6
post +belief_sim +traits_feat_r	60.35	63.07	0.63	0.6	0.6
traits_feat_p +post +traits_feat_r +traits_feat_b +belief_sc	59.71	62.47	0.63	0.6	0.6
traits_feat_p +post +traits_flat +traits_feat_r +belief_sc	59.51	62.3	0.63	0.59	0.59
traits_feat_p +post +traits_flat +traits_feat_r +traits_feat_b +belief_sc	59.51	62.3	0.63	0.59	0.59
traits_feat_p +post +traits_feat_r +belief_sc	59.62	62.39	0.63	0.6	0.59
traits_feat_p +post +belief_sim +traits_feat_r +traits_feat_b +belief_sc	59.79	62.46	0.63	0.6	0.6
traits_feat_p +post +traits_flat +belief_sim +traits_feat_r +traits_feat_b +belief_sc	59.82	62.58	0.63	0.6	0.6
traits_feat_p +post +traits_flat +belief_sim +traits_feat_r +belief_sc	59.82	62.58	0.63	0.6	0.6
traits_feat_p +post +belief_sim +traits_feat_r +belief_sc	59.79	62.46	0.63	0.6	0.6
traits_feat_p +post +belief_sc	59.68	62.5	0.63	0.6	0.59
traits_feat_p +post +traits_feat_b +belief_sc	59.65	62.46	0.63	0.6	0.59
traits_feat_p +post +traits_flat +belief_sc	59.71	62.47	0.63	0.6	0.6
traits_feat_p +post +traits_flat +traits_feat_b +belief_sc	59.74	62.51	0.63	0.6	0.6
traits_feat_p +post +belief_sim +belief_sc	59.31	62.05	0.63	0.59	0.59
traits_feat_p +post +belief_sim +traits_feat_b +belief_sc	59.31	62.07	0.63	0.59	0.59
traits_feat_p +post +traits_flat +belief_sim +traits_feat_b +belief_sc	59.37	62.08	0.63	0.59	0.59

Bibliography

Robin M Akert and Abigail T Panter. Extraversion and the ability to decode nonverbal communication. *Personality and Individual Differences*, 9(6):965–972, 1988. 4.2

Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001. 2.3

Antoine C. Braet. Ethos, pathos and logos in aristotle's rhetoric: A re-examination. *Argumentation*, 6(3):307–320, Aug 1992. ISSN 1572-8374. doi: 10.1007/BF00154696. URL `https://doi.org/10.1007/BF00154696`. 2.2

Keith Coaley. The assessment and measurement of personality. In *An Introduction to Psychological Assessment and Psychometrics*, pages 177–204, London, 2010. SAGE Publications Ltd. doi: 10.4135/9781446221556.n8. URL `http://dx.doi.org/10.4135/9781446221556.n8`. 2.2.1

Esin Durmus and Claire Cardie. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1094. URL `http://aclweb.org/anthology/N18-1094`. 1, 2.4

B. J. Fogg. Mass interpersonal persuasion: An early view of a new phenomenon. In Harri Oinas-Kukkonen, Per Hasle, Marja Harjumaa, Katarina Segerståhl, and Peter Øhrstrøm, editors, *Persuasive Technology*, pages 23–34, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-68504-3. 1

Lewis R. Goldberg. An alternative" description of personality": The big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990. 2.2.1

Ivan Habernal and Iryna Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany, August

2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1150. URL https://www.aclweb.org/anthology/P16-1150. 1

Ivan Habernal and Iryna Gurevych. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas, November 2016b. Association for Computational Linguistics. doi: 10.18653/v1/D16-1129. URL https://www.aclweb.org/anthology/D16-1129. 1

Jos Hornikx. A review of experimental research on the relative persuasiveness of anecdotal, statistical, causal, and expert evidence. *Studies in Communication Sciences*, 5(1):205–216, 2005. 1

C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text, 2014. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109. 2.3

Bing Liu and Lei Zhang. *A Survey of Opinion Mining and Sentiment Analysis*, pages 415–463. Springer US, Boston, MA, 2012. ISBN 978-1-4614-3223-4. doi: 10.1007/978-1-4614-3223-4_13. URL https://doi.org/10.1007/978-1-4614-3223-4_13. 2.3

Stephanie M Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. *arXiv preprint arXiv:1708.09085*, 2017. 1

Olena Medelyan, Ian H Witten, and David Milne. Topic indexing with wikipedia. In *Proceedings of the AAAI WikiAI workshop*, volume 1, pages 19–24, 2008. 3.1.2

Rada Mihalcea and Andras Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. doi: 10.1145/1321440.1321475. URL http://doi.acm.org/10.1145/1321440.1321475. 3.1.2

David Milne and Ian H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458150. URL http://doi.acm.org/10.1145/1458082.1458150. 2.3, 3.1.2

Daan Odijk, Edgar Meij, and Maarten de Rijke. Feeding the second screen: Semantic linking based on subtitles. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 9–16, Paris, France, France, 2013. LE CENTRE DE HAUTES ETUDES INTERNATIONALES

D'INFORMATIQUE DOCUMENTAIRE. ISBN 978-2-905450-09-8. URL http://dl.acm.org/citation.cfm?id=2491748.2491751. 3.1.2

Daniel J O'Keefe. Standpoint explicitness and persuasive effect: A meta-analytic review of the effects of varying conclusion articulation in persuasive messages. *Argumentation and Advocacy*, 34(1):1–12, 1997. 3.1.5

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001. 2.2.1, 3.1.3

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003. 3.1.3

Richard E Petty. *Attitudes and persuasion: Classic and contemporary approaches*. Routledge, 2018. 1

Chanthika Pornpitakpan. The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, 34(2):243–281, Feb 2004. doi: 10.1111/j.1559-1816.2004.tb02547.x. URL http://dx.doi.org/10.1111/j.1559-1816.2004.tb02547.x. 1

Jonathan Raphael Raiman and Olivier Michel Raiman. Deeptype: multilingual entity linking by neural type system evolution. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2.3

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013. 3.1.3

Brian Sternthal, Ruby Dholakia, and Clark Leavitt. The persuasive effect of source credibility: Tests of cognitive response. *Journal of Consumer Research*, 4(4): 252–260, 1978. ISSN 00935301, 15375277. URL http://www.jstor.org/stable/2488816. 1

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 613–624, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1. doi: 10.1145/2872427.2883081. URL https://doi.org/10.1145/2872427.2883081. 1, 2.1.2, 2.1.2, 2.4, 3, 3.1.5, 3.1.5, 3.2, 3.2, 4.2, 4.2, 4.2

Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010a. doi: 10.1177/0261927X09351676. 2.3

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010b. 3.1.3

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010. 3.1.3

Ernest C. Tupes and Raymond E. Christal. Recurrent personality factors based on trait ratings. *Journal of Personality*, 60(2):225–251, Jun 1992. doi: 10.1111/j.1467-6494.1992.tb00973.x. URL `http://dx.doi.org/10.1111/j.1467-6494.1992.tb00973.x`. 2.2.1

Bonachristus Umeogu. Source credibility: A philosophical analysis. *Open Journal of Philosophy*, 2(2):112–115, 2012. doi: 10.4236/ojpp.2012.22017. URL `http://dx.doi.org/10.4236/ojpp.2012.22017`. 1

Tal Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373, Jun 2010. doi: 10.1016/j.jrp.2010.04.001. URL `http://dx.doi.org/10.1016/j.jrp.2010.04.001`. 2.3