The University of Sheffield

A Computational Study of Speech Acts in Social Media

A thesis submitted for the degree of Doctor of Philosophy

in the

Department of Computer Science

Mali Jin

Supervisor: Prof. Nikolaos Aletras

December 2022

Acknowledgments

I would first like to thank my supervisor, Nikolaos Aletras, for his invaluable guidance, patience and understanding throughout my studies. Nikos supported me to the fullest not only academically and professionally but also personally, which allowed me to develop and enjoy my PhD studies.

I would also like to thank the panel committee members, Elif Bilge Kavun, Loic Barrault, and my second supervisor, Chenghua Lin, for their invaluable advice at my early study stages.

My gratitude also goes to my collaborators in research projects about studying the speech act of bragging, A. Seza Doğruöz and Daniel Preoţiuc-Pietro. Their patient guidance, overall insights and knowledge in this field have made this an inspiring experience for me.

Also, I would like to express my thanks to my colleagues from my late study stages as a part-time research associate at the University of Sheffield, Kalina Bontcheva, Xingyi Song, Diana Maynard and Johann Petrak. It is an honor to have a chance to participate in exciting projects and work with such brilliant researchers.

Many thanks to all my colleagues in Natural Language Processing group for making a great working environment. I really appreciated helpful discussions and after-work activities. They have made these years of my studies very enjoyable.

Finally, I would like to thank my family and friends for their support and encouragement. I especially would like to express my deepest thanks to my parents. Their unreserved love and support give me the courage to face everything.

Abstract

Speech acts are expressed by humans in daily communication that perform an action (e.g. requesting, suggesting, promising, apologizing). Modeling speech acts is important for improving natural language understanding (i.e. human-computer interaction through computers' comprehension of human language) and developing other natural language processing (NLP) tasks such as question answering and machine translation. Analyzing speech acts on large scale using computational methods could benefit linguists and social scientists in getting insights into human language and behavior.

Speech acts such as suggesting, questioning and irony have aroused great attention in previous NLP research. However, two common speech acts, complaining and bragging, have remained under explored. Complaints are used to express a mismatch between reality and expectations towards an entity or event. Previous research has only focused on binary complaint identification (i.e. whether a social media post contains a complaint or not) using traditional machine learning models with feature engineering. Bragging is one of the most common ways of self-presentation, which aims to create a favorable image by disclosing positive statements about speakers or their in-group. Previous studies on bragging have been limited to manual analyses of small data sets, e.g. fewer than 300 posts.

The main aim of this thesis is to enrich the study of speech acts in computational linguistics. First, we introduce the task of classifying complaint severity levels and propose a method for injecting external linguistic information into novel pretrained neural language models (e.g. BERT). We show that incorporating linguistic features is beneficial to complaint severity classification. We also improve the performance of binary complaint prediction with the help of complaint severity information in multi-task learning settings (i.e. jointly model these two tasks). Second, we introduce the task of identifying bragging and classifying their types as well as a new annotated data set. We analyze linguistic patterns of bragging and their types and present error analysis to identify model limitations. Finally, we examine the relationship between online bragging and a range of common socio-demographic factors including gender, age, education, income and popularity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The “speech act theory” originates from the field of linguistics and psychology (Austin, 1962; Searle, 1969). It considers language as an action, which states that when people say something, they actually do something about it (Austin, 1962). Speech acts are utterances used by individuals to perform an action such as making statements, asking questions, requesting, apologizing, thanking, inviting, warning and congratulating (Searle, 1969). For example, *“Can you close the door please?”* expresses the speaker’s request for someone to close the door. Speech acts are common means used in daily communication to convey speakers’ attitudes, intentions and behaviors.

In recent years, there has been a surge of interest in automatically detecting speech acts using NLP technologies (Wang and Chua, 2010; Karoui et al., 2015; Negi et al., 2019; Anikina and Kruijff-Korbayova, 2019; Saha et al., 2020). However, due to their implicit or ambiguous expressions, automatic identification of speech acts becomes a challenging task (Anikina and Kruijff-Korbayova, 2019; Saha et al., 2020). Moreover, typos, emojis, improper grammar and informal terms in social media make it more difficult to analyze speech acts online. Table 1.1 shows examples of supervised machine learning models (i.e. algorithms that are trained on labeled data set for prediction) failing to identify an actual speech act in a social media text (i.e. tweet) when a contextual understanding is required (Jin and Aletras, 2021; Jin et al., 2022). In the case of the first tweet, the key factor of identifying it as a complaint is that the speaker is not satisfied with how the exhaust has been repaired, yet the model cannot extract such information.

Modeling speech acts is crucial to natural language understanding (NLU), which enables

Tweet	**True**	**Prediction**
Is this how you fix the exhaust of your <USER> in #belarus? <URL>	Complaint	Non-complaint
9 hr drives feel like nothing now lol	Bragging	Non-bragging

Table 1.1: Examples of a model failing to detect a complaint and bragging.

computers to communicate with people in natural language by reading and comprehending texts (Schank, 1972). It is beneficial in various NLP tasks such as question answering (Simmons, 1965) and machine translation (Brown et al., 1990). It can also provide linguistic insights (e.g. actions, attitudes or intentions) in a text which may not be easy to extract with traditional sentiment analysis models (Preoţiuc-Pietro et al., 2019). Furthermore, studying speech acts in computational linguistics can help linguists and social scientists to better understand humans and language (Dayter, 2014; Van Damme et al., 2017; Sezer et al., 2018).

Previous work in NLP has focused on studying various speech acts such as suggesting (Negi et al., 2019; Anand et al., 2019), questioning (Wang and Chua, 2010; Prabowo and Herwanto, 2019) and irony (Karoui et al., 2015; Van Hee et al., 2018). However, two common speech acts, complaining and bragging, have yet to receive significant attention.

Complaining expresses a negative mismatch between reality and expectations towards an entity or event (Olshtain and Weinbach, 1987) (e.g. *"Terrible service! They kept me waiting for 2 hours in the store!"*). Complaints appear frequently on online social networks, especially in customer reviews (Vásquez, 2011). They are used to convey special demands of speakers such as venting negative emotions, requiring apologies or seeking solutions (Kowalski, 1996). Complaining is regarded as face-threatening acts (Brown and Levinson, 1987) as they may damage the face (i.e. public image) or self-esteem of the recipient who is responsible for this act (Goffman, 1967). The ability to automatically identify complaints and their severity levels (e.g. hint without directly mentioning the dissatisfaction, blame complainees directly for their action) is vital for understanding users' needs and improving customer service (Au et al., 2009; Vásquez, 2011).

The speech act of bragging aims to construct a favorable self-image by disclosing positive statements about the speaker or their in-group (Dayter, 2014, 2018) (e.g. *"Finally got the offer! Whoop!!"*). Previous work in pragmatics has shown that bagging is found to be more frequent in social media than in face-to-face interactions (Ren and Guo, 2020). It is one of the most common strategies used for online self-presentation to meet speakers' goals such as gaining popularity in certain communities (Dayter, 2018). Although bragging online is predominantly positive, it is also considered a high-risk act as it threats the positive face (i.e. the desire to be liked) under politeness theory (Brown and Levinson, 1987). Inappropriate bragging may lead to the opposite effect than intended such as dislike. Studying bragging in computational linguistics is helpful for online users to enhance their self-presentation strategies (Miller et al., 1992; Dayter, 2018).

1.1 Aims and Objectives

This thesis focuses on modeling complaining and bragging in social media using methods from computational linguistics and machine learning. With the underlying challenges, it aims to achieve the following research objectives:

- Previous work on modeling complaining has focused on distinguishing complaints from non-complaints in different domains (Coussement and Van den Poel, 2008; Jin et al., 2013; Preoţiuc-Pietro et al., 2019) or classifying them based on task-specific scenarios such as responsible departments (Laksana and Purwarianti, 2014; Tjandra et al., 2015; Gunawan et al., 2018), possible hazards and risks (Bhat and Culotta, 2017) and escalation likelihood in customer service (Yang et al., 2019a). First, we aim to enrich the complaint classification tasks by recognizing dissatisfaction levels and intentions of the complainer with fine-grained categories.

- Previous work on automatically identifying complaints in social media has focused on using feature-based and task-specific neural network models (Coussement and Van den Poel, 2008; Jin et al., 2013; Preoţiuc-Pietro et al., 2019). However, there is a need for further efforts to improve the complaint classification models by using state-of-the-art text encoding methods based on pretrained language models.

- The speech act of bragging has been extensively studied in pragmatics and psychology across different languages (Dayter, 2014; Scopelliti et al., 2015; Matley, 2018; Ren and Guo, 2020). However, bragging has yet to be studied at scale in computational linguistics. Therefore, we aim to develop new data resources and models for studying bragging on a large scale.

- Previous work in computational sociolinguistics and computational social science showed that user traits (e.g. age, gender and personality) correlate with language use and online behavior (Nguyen et al., 2016). User features and temporal clues have been used to study Facebook language (Schwartz et al., 2013), user income (Preoţiuc-Pietro et al., 2015b), sentiment analysis and topic classification (Hovy, 2015), hate speech detection (Fehn Unsvåg and Gambäck, 2018) and suicidal ideation detection (Cao et al., 2020; Sawhney et al., 2020). By using user traits and temporal patterns, we aim to examine individual and temporal differences in bragging and to investigate who brags, when and how they do it in online environments.

1.2 Contributions

The main contributions made throughout this thesis are as follows:

- We enrich a publicly available data set of complaints with four severity categories based on the linguistic theory of pragmatics.
- We propose multi-task learning (MTL) architectures jointly modeling complaint severity classification and complaint identification, which achieve state-of-the-art results on complaint identification.
- We introduce a new publicly available data set annotated with bragging and their types.
- We propose an approach that introduces linguistic information into transformer networks to improve the performance of complaint and bragging prediction.
- We present the first large-scale study of bragging and its correlation with user socio-demographic factors in computational linguistics.

1.3 Thesis Overview

Chapter 2 begins with the background knowledge of speech acts in pragmatics and focuses on two types of speech acts mainly involved in this thesis, complaining and bragging. This includes their definitions, their impact on social media, the way they are expressed and related linguistic analysis. It then describes speech act detection as a text classification task in NLP. Finally, it discusses the computational approaches used in modeling speech acts from the previous NLP work.

Chapter 3 introduces a new classification task for complaint severity identification. It enriches an existing complaint data set with four different severity levels based on linguistic theory. It proposes an approach that injects external linguistic information into transformer networks. Results show that the proposed methods perform better than the vanilla model in this task.

Chapter 4 evaluates various transformer-based models and their combinations with linguistic features for complaint identification (i.e. identifying whether a text is a complaint or not). It also describes MTL architectures that jointly model complaint identification and complaint severity classification. Results demonstrate that the proposed MTL settings outperform state-of-the-art methods for complaint identification.

Chapter 5 first motivates the importance of studying bragging using computational approaches. It introduces a new data set annotated with bragging and its types. The data set has been made publicly available. It then presents the computational study of bragging based on two new tasks using the data set: bragging identification and bragging type classification. Similar to the complaint classification task, it explores transformer-based models combined with different linguistic features. Results as well as linguistic analysis reveal markers of bragging in tweets and model behavior in predicting bragging.

Chapter 6 presents a large scale study of bragging behavior by U.S. Twitter users. It makes use of a state-of-the-art predictive model for bragging identification based on their tweets. The data set used to conduct the analysis is constructed from a group of 2,685 users over 10 years for which we have access to self-reported demographic traits. It introduces an approach to normalizing bragging percent in order to account for a temporal effect in the proportion of bragging. Then it examines individual and temporal differences in online bragging among these users based on their bragging ratios.

Chapter 7 summarises the findings of this thesis and suggests possible directions for future research on these topics.

1.4 Publications

Work contributing to this thesis has been published in the following peer-reviewed venues:

- The work presented in Chapter 3 has been published at the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021) (Jin and Aletras, 2021);
- The work presented in Chapter 4 has been published at the Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020) (Jin and Aletras, 2020) and the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021) (Jin and Aletras, 2021);
- The work presented in Chapter 5 has been published at the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) (Jin et al., 2022);
- The work presented in Chapter 6 has been submitted to EPJ Data Science (under review).

Chapter 2

Background

The aim of this chapter is to provide a background on speech acts in pragmatics and how these have been analyzed using NLP techniques. Furthermore, this chapter identifies and discusses the limitations of previous work in computational pragmatics.

2.1 Speech Acts in Pragmatics

Speech acts stem from the theory that language is the medium for performing social acts (Austin, 1962). This means that language is used not only to express things but also to do things. This theory is often used in the field of philosophy of language and linguistics. Austin (1962) defined three levels of actions performed by language in parallel: locutionary, illocutionary and perlocutionary. Firstly, the locutionary act is the act of saying something, which consists of the actual utterance and the understandable meaning that it conveys or expresses. Secondly, the illocutionary act contains the actual intention of the utterance (i.e. its semantic force), which might be implied by certain tones, attitudes, feelings or emotions. Finally, the perlocutionary act refers to the consequential effects on the audience, which are usually in the form of emotions (e.g. threat, warning), feelings (e.g. encouragement) or thoughts (e.g. suggestion). For example, if the locutionary act describes a dangerous situation, the illocutionary act indicates a warning intent and the perlocutionary act might frighten the audience.[1]

[1] The example is from https://www.communicationtheory.org/speech-act-theory/#:~:text=The%20speech%20act%20theory%20considers,How%20do%20things%20with%20words'.

Later, Searle (1969) classified the speech acts into five main categories based on the function indicated in the illocutionary act: declarations, representatives, commissives, directives and expressives. First, declarations are used by the speaker to change the state of affairs (e.g. resigning, appointing); second, representatives are used to denote what the speaker believes to be the matter (e.g. claiming, swearing); third, commissives are used to perpetrate the speaker to carry out a future action (e.g. promising, rejecting); then, directives are used to require the hearer to take a particular action (e.g. requesting, inviting); finally, expressives are used to state the speaker's feelings or attitudes (e.g. apologizing, complaining).

Speech acts are commonly used by individuals in daily communication to perform an action, which usually expect the audience to react with verbal or non-verbal behavior (Austin, 1962). There are many types of speech acts, which represent different functions, such as requesting (ask the audience to do something without it being obvious that he/she would perform the action in the normal course of events; Searle 1969), suggesting (get the audience to commit him/herself to some future course of action; Searle 1976), promising (commit to future actions for the benefit of the audience; Searle 1969), apologizing (express regret over offensive acts; Válková 2013). However, this thesis focuses on complaining and bragging where there is little prior work in NLP.

2.2 Analysis of Complaining in Linguistics and Psychology

Complaining is a basic speech act expressing special demands of speakers. They are made in order to vent negative emotions or reach a certain goal (Kowalski, 1996) such as apologies or reparations. These negative emotions are triggered by a discrepancy between reality and expectations towards an entity or event (Olshtain and Weinbach, 1987). Taking customer reviews as an example, complaints are likely to be made when customers receive a product that is inconsistent with its description. In this situation, the remedy can be an exchange or refund. Table 2.1 shows three reviews from social media, where the difference between a complaint and a negative comment is whether it expresses a breach of expectations (e.g. the delivery arrived later than the speaker expected).

Complaining is commonly used in daily communication and has attracted extensive attention in linguistics (Olshtain and Weinbach, 1987; Sacks, 1992; Boxer, 1993a). However,

Review type	**Example**
Neutral review	*How long does ur standard shipping take?*
Negative review	*The browns are the worst*
Complaint	*How come the weekend delivery arrives so late in the morning - I'd love to read w / my coffee!*

Table 2.1: Examples of a neutral review, negative review and complaint.

due to the variety of expressions in complaints, it is difficult for researchers to clearly define them. On one hand, there are no typical words or phrases representing complaints (Wolfe and Powell, 2006). For example, complaints cannot be expressed in a fixed structure, unlike questions that usually begin with the word "wh". On the other hand, complaints often co-occur with other speech acts such as suggestions, criticism, admonitions and threats (Wolfe and Powell, 2006), which makes them more complicated (e.g. *"The service and taste in this restaurant do not match the price. You'd better not go!"*).

There are multiple definitions for complaining as a speech act. It has been defined as a piece of praise plus "but" plus something else which is usually an expression of dissatisfaction (Sacks, 1992). Similarly, Kowalski (1996) states that a complaint is a kind of comment with dissatisfaction; while Heinemann and Traverso (2009) extend it to nearly any comment with the slightest negative stance. According to a more specific explanation proposed by Olshtain and Weinbach (1987), a complaint happens when speakers expect a favorable event to occur or an unfavorable event to be avoided, but their expectations are violated. We adopt the definition of Olshtain and Weinbach (1987) in this thesis to be consistent with the previous studies of complaints in computational linguistics (Preoţiuc-Pietro et al., 2019).

By complaining, the complainant can vent their negative emotions in a timely manner or make the problem properly solved. However, it is regarded as a high-risk act (Brown and Levinson, 1987) as it may threaten the face of the addressee, which represents the public image of a person or an entity (Goffman, 1967). According to the theory of Goffman (1967), there are two aspects of the face: positive (i.e. the desire to be liked) and negative face (i.e., the desire not to be imposed). Complaining is likely to damage both positive and negative faces. On one hand, positive faces can be influenced by dissatisfaction with the responsible party, which results from the destruction of expectations. On the other hand, complaints are usually made with the intent to request a remedy or compensation. Therefore, negative faces are affected by forcing the responsible party to take action.

Complaint type	**Example**
Direct complaint	*Whoever owns it now really should get out of the hotel business forever!!!!!! It is obvious they have no clue.*
Indirect complaint	*If you are at all of a delicate constitution with regards cleanliness, stay at some other hotel.*
Both direct and Indirect complaint	*Who ever owns this should be ashamed, this is not a 3 star it's a no star. [...]Do yourself a big favour and give this place complete miss.*

Table 2.2: Examples of direct and indirect complaints and their combination (Vásquez, 2011).

2.2.1 Types of Complaints

The way a complaint is expressed varies from person to person or from situation to situation. Different types of complaints are defined according to different criteria in linguistics such as the recipient of complaints and the severity of expressions. There are direct and indirect complaints depending on whether or not the recipient is the responsible party (Sacks, 1992). Also, complaints can be divided into different severity levels based on their explicitness and speakers' purpose (Olshtain and Weinbach, 1987; Trosborg, 2011).

Direct versus indirect As defined by Boxer (1993a), a direct complaint refers to a complaint that is addressed to the party who is responsible for violating the expectation of the speakers; while an indirect complaint expresses the speakers' dissatisfaction with themselves or someone who is not present. In the case of an indirect complaint, the recipient is a third party who does not take responsibility or does not need to make compensation for the complainer (e.g. other users online). Vásquez (2011) found that most of the direct complaints usually contain a third person pronoun such as 'whoever' and 'they'; while a second person pronoun, or an imperative sentence, or both, often appears in indirect complaints, which intend to give suggestions, instructions or warnings to other online users. Table 2.2 shows examples of a direct and indirect complaint and their combination, where the direct one is against the hotel owner and the indirect one aims at all users who might see it.

Severity level In pragmatics, complaints have been classified into different levels of severity according to their directness, the amount of face-threat that the complainer is willing to

Complaint severity level	**Example**
Below the level of reproach	*No harm done, let's meet some other time.*
Expression of annoyance of annoyance or disapproval	*It's a shame that we have to work faster now.*
Explicit complaint	*You are always late and now we have less time to do the job.*
Accusation	*Next time don't expect me to sit here waiting for you.*
Warning, immediate threat	*If we don't finish the job today I'll have to discuss it with the boss.*

Table 2.3: Five categories of complaints severity level based on a specific scenario (Olshtain and Weinbach, 1987).

undertake and their purpose (e.g. to vent dissatisfaction or to look for solutions).

Olshtain and Weinbach (1987) divided complaints into 5 distinct categories: (a) below the level of reproach; (b) expression of annoyance or disapproval; (c) explicit complaint; (d) accusation; (e) warning, immediate threat. Table 2.3 presents examples of each level based on a specific scenario where one colleague had waited for another one.

More recently, Trosborg (2011) defined four major severity levels: (a) no explicit reproach; (b) disapproval; (c) accusation; (d) blame. "No explicit reproach" means there is no offense in the statement while "disapproval" expresses speakers' negative emotions only (e.g. dissatisfaction, annoyance, dislike or disapproval) without mentioning complainees. The difference between "accusation" and "blame" is that the latter one emphasizes the responsibility of the person being complained about. Table 2.4 presents the original definitions and examples of each category as well as their sub-categories by Trosborg (2011) (more details will be explained in Chapter 3).

Finally, Kakolaki and Shahrokhi (2016) classified complaints into levels of directness: (a) very direct; (b) somewhat direct; (c) indirect. Notably, "direct" and "indirect" here refer to explicit and implicit, which is different from the definition of Boxer (1993a) in *Direct versus indirect*. Direct complaints (i.e. very direct and somewhat direct) include obvious breaches of expectations; while indirect complaints do not explicitly mention or imply a breach of expectations. Moreover, the difference between very direct and somewhat direct is that the former highlights the responsibility of the complaint recipient while the latter does not.

Severity Level	**Definition**	**Example**
I No explicit reproach 1 Hints	Complainer does not mention the complaint in the complainable and does not directly state something is offensive.	*My car was in perfect order when I last drove it. There was nothing wrong with my car yesterday.*
II Disapproval 2 Annoyance 3 Consequences	Complainer expresses dislike, disapproval, and annoyance in connection with a certain state of affairs that he or she considers bad for him or her.	*There's a horrible dent in my car. Oh dear, I've just bought it.* *How terrible! Now I won't be able to get to work tomorrow. Oh, damn it, I'll lose my insurance bonus now.*
III Accusation 4 Indirect 5 Direct	Complainer establishes the complainee as the agent of the complainable and directly or indirectly accuses the complainee for committing the problem.	*You borrowed my car last night, didn't you?* *Did you happen to bump into my car?*
IV Blame 6 Modified blame 7 Explicit blame (behavior) 8 Explicit blame (person)	Complainer assumes that the complainee is guilty of the offence and states modified blame of complainee's action or directly blames the complainee or his or her action.	*Honestly, couldn't you have been more careful?* *You should take more care with other people's car.* *It's really too bad, you know, going round wrecking other people's car. How on earth did you manage to be so stupid?* *Oh no, not again! You are really thoughtless.* *Bloody fool! You've done it again!*

Table 2.4: Four major categories of complaints severity level by Trosborg (2011).

2.2.2 Qualitative Studies on Complaining

The speech act of complaining is used by complainers to express their demands due to the breach of their expectations. Complaining is also considered a high-risk act as it may verbally attack the addressee who is responsible for the unpleasant result. Therefore, strategies (e.g. politeness and indirect strategy) are usually applied in complaint utterances in order to mitigate the offense and avoid being impolite, rude or disrespectful (Wannaruk, 2008). Different uses of pragmatic strategies cause different expressions of complaints to some extent. The phenomenon of making complaints under various complaint strategies has attracted the attention of researchers. Studies have observed that the degree of politeness and directness in complaint expressions is related to environmental, individual and cultural background (Boxer, 1993a; Moon, 2001; Geluykens and Kraft, 2007; Vásquez, 2011; Ghahraman and

Nakhle, 2013).

Complaints develop differently in online environments where people are more likely not to know each other, which is contrary to face-to-face communications. In face-to-face interactions, complaints tend to be expressed in an implicit and euphemistic way as the speakers are more vulnerable under exposure (Heinemann and Traverso, 2009). Conversely, complainers may make more direct and explicit complaints online without concerning about being identified (Vásquez, 2011). Also, online complaints tend to co-occur more frequently with advice and recommendations than face-to-face complaints as the potential audiences might be other online users instead of the actual complainee (Vásquez, 2011).

The social distance between speakers and complainees and social status is one of the factors that influence the way speakers complain. Boxer (1993b) investigated the impact of social distance (e.g. intimates, friends, acquaintances, strangers) on complaining behaviors. People tend to behave differently with intimates than with friends and strangers (Boxer, 1993a).

Noisiri (2002) examined the frequency of complaint strategies used by Thai native speakers according to different social distances. When the addressee is a waiter, the highest level of complaint severity (i.e. "blame") is used more frequently; when the addressee is a flatmate, they tend to use a strong complaint statement, but it is less aggressive than in the first situation; when the addressee is an elder stranger, women prefer using the lowest level of complaint severity (i.e. "hint") while men use "blame" more frequently.

Kaharuddin (2020) compared the complaint strategies used by English native speakers (ENSs) and Indonesian native speakers (INSs) on friends, intimates (e.g. family members) and strangers (e.g. car drivers) through a scenario-based questionnaire. The results showed that both ENSs and INSs make implicit complaints to their friends and make explicit complaints to strangers; while to intimates, INSs express complaints in an explicit way more frequently than ENSs do.

Al-Shboul (2021) investigated the influence of social status and social distance on the production of complaints by Jordanian students. According to the study, they make complaints in a less direct way to people with a higher social status (e.g. professors). Also, more direct strategies are likely to be used to friends (e.g. classmates) than to strangers (e.g. service people).

Previous linguistic research has analyzed complaint behavior between native English speakers and non-native English speakers (Neu, 1996; Moon, 2001; Tanck, 2002). Due to a lack of knowledge of the second language, it is difficult for non-native speakers to complain in English (Noisiri, 2002). Thus, they are more likely to complain in an inappropriate way (e.g. confrontational, presumptuous, vague) and tend to make explicit complaints while native speakers use implicit complaints mostly (Moon, 2001; Tanck, 2002).

Also, various studies have examined gender differences in choosing complaint strategies among different groups such as Thai native speakers (Noisiri, 2002), American students (Wolfe and Powell, 2006), Indonesian English as a foreign language (EFL) students (Sukyadi et al., 2011), Canadian native speakers (Ghahraman and Nakhle, 2013), Iranian EFL students (Kakolaki and Shahrokhi, 2016) and Jordanian students (Al-Shboul, 2021). These studies concluded that males tend to make direct and aggressive complaints more frequently whereas females usually express them in an indirect and soft manner way. This can be explained by the findings that the speech of women is more polite than that of men and women show more interest in building friendships while men prefer being more independent (Holmes, 2013). Additionally, Wolfe and Powell (2006) analyzed the reasons why men and women complain and found that women are more likely to use complaints as an indirect request to influence others' actions while men make complaints to excuse their behaviors.

Furthermore, some studies have focused on complaints in online reviews (Au et al., 2009; Maurer and Schaich, 2011; Vásquez, 2011). Analyzing complaint information in customer reviews is beneficial to gain a better understanding of customers' needs and preferences and thus improve customer service and marketing strategies. For example, Au et al. (2009) conducted an analysis of individual complaint cases reported on TripAdvisor for Hongkong hotels and observed that customers tend to care more about some fundamental services (e.g. service delivery and employee behavior). Also, issues such as slow responses and poor compensation were also identified when dealing with online complaints in the studies (Au et al., 2009).

2.3 Analysis of Bragging in Linguistics and Psychology

The desire to be viewed positively is a key driver of human behavior (Baumeister, 1982; Leary and Kowalski, 1990; Sedikides, 1993; Tetlock, 2002) and creating a positive image often leads to personal rewards (Gilmore and Ferris, 1989; Hogan, 1982; Schlenker, 1980). Self-presentation strategies are means for individuals to build and establish a positive social

Type	**Example**
Bragging	*Just impressed myself with how much French I think I undersood! One semester at KC FTW!*
Non-bragging	*Glad to hear that! Well done Jim!*

Table 2.5: Examples of bragging and non-bragging.

image to meet their goals (e.g. gaining popularity in certain communities) (Goffman et al., 1978; Jones et al., 1982; Jones, 1990; Bak et al., 2014). Bragging (or self-praise) is one of the most common strategies and involves uttering a positive statement about oneself or their close networks such as team members or family members (Dayter, 2014). As defined by Dayter (2014), it is *a speech act which explicitly or implicitly attributes credit to the speaker for some "good" which is positively valued by the speaker and the potential audience.* Bragging content is usually everyday achievements or personal qualities (Matley, 2018). Table 2.5 shows examples of bragging and non-bragging. Despite the positive sentiment in both texts, the first one discloses the speaker's ability to learn French toward which the speaker takes a positive stance.

Bragging as a speech act is considered a face-threatening act to positive faces (i.e. the desire to be liked) under politeness theory (Brown and Levinson, 1987). It is directly oriented to the speaker and may threaten their likeability if the bragging is perceived negatively, while also may affect listeners' faces by implying that their feelings are not valued by the speaker (Matley, 2018).

2.3.1 Strategies of Bragging

Modest and sincere self-presentation styles are more likely to be perceived positively (Sedikides et al., 2007). Bragging framed as mere information-sharing, but with positive connotation to the speaker, can make the speaker be perceived as more likeable (Miller et al., 1992). Researchers suggest that merely sharing information or involving statements of achievement in a modesty form is more likely to be flattering (Miller et al., 1992; Matley, 2018). However, it can be perceived negatively and causes greater aggression when it involves boasting, focusing on the nature of the person (e.g. *"I'm a wonderful person"*), elements of competitiveness (e.g. *"I am better than you"*), use of superlatives (e.g. *"I was the best player"*) and explicit comparisons to others (Miller et al., 1992; Hoorens et al., 2012; Scopelliti et al., 2015; Matley, 2018). In addition, competence-related statements are more likely to be negatively perceived than those based on warmth (e.g. the ability to form connections with others) (Van Damme et al., 2017).

As speakers appear to be aware of the potential negative effects of bragging, mitigation strategies are usually applied in bragging utterances. Bragging can be classified according to whether and what kind of strategy is used. Dayter (2014) and Ren and Guo (2020) defined two types of explicit bragging: *explicit bragging without modification* and *explicit bragging with modification*. Explicit bragging without modification expresses something good valued by the speaker straightforwardly without embellishment or cover-up; while explicit bragging with modification contains a bragging statement plus something (e.g. disclaimer, shift of focus, self-denigration and reference to hard work) to attenuate praising themselves. Dayter (2014) defined a third category named *reinterpretation*, where a bragging statement is followed by a complaint or reshaped into a complaint. Later, Ren and Guo (2020) expanded the types of speech acts that bragging disguises (e.g. bragging as a complaint, bragging as a question, a narration and sharing) and name them *implicit bragging* where the exterior meaning of the utterance is different from the illocutionary act.

The following common strategies were identified in the studies (Dayter, 2014; Sezer et al., 2018; Matley, 2018; Ren and Guo, 2020; Maíz-Arévalo, 2021), which help speakers to mitigate the social risk and negative effects caused by bragging (Scopelliti et al., 2015; Tobback, 2019a; Matley, 2020):

- **Shifting the focus.** Speakers shift the credit when shifting the praise focus from themselves to a person who is closely related to them, which is a safer way than directly praising themselves (Ren and Guo, 2020). E.g. *"My son came first in the 50-meter run. Like father, like son."*

- **Reference to hard work.** Speakers attribute their achievements to their efforts and the greater their effort, the more likely they are to be perceived favorably (Miller et al., 1992). E.g. *"After a month of practice, I finally succeeded in becoming a band drummer in our school!"*

- **Reporting bragging.** Speakers reframe bragging as praise from a third party rather than their own, which is helpful to mitigate the face threat as the praise has been verified by others (Ren and Guo, 2020). E.g. *"My colleagues always say I make the best coffee."*

- **Disclaimers.** Disclaimers is a discursive device to preemptively avoid or defeat doubts and negative perceptions that may arise from the intended act (Hewitt and Stokes, 1975). Disclaimers are employed as a remedial measure to minimize bragging. This includes admitting it is not right to brag (e.g. *"it is wrong for me to brag but..."*), apologizing (e.g. *"Sorry if you already saw it..."*) and denying compliments (e.g. *"Not to brag but..."*) (Dayter, 2014; Matley, 2018; Maíz-Arévalo, 2021).

- **Aggravation.** Speakers openly express their intent to brag by admitting the self-praise behavior (e.g. using hashtags such as #brag and #humblebrag) (Matley, 2018). In this way, the intended illocutionary act is apparent so that audiences clearly know the speaker is bragging. E.g. *"I want to #brag that I passed the exam."*

- **Collectivism.** Speakers brag about the collective as a member of the collective by using "we" or "our" instead of "I" (Ren and Guo, 2020). The aim of this strategy is to shift the praise focus from an individual to a group of people. E.g. *"Our team did a great job in this project. We are the best!"*

- **Self-denigration.** Speakers depreciate themselves before bragging. They attempt to enhance their face immediately after self-denigration in order to restore the balance and reassure the audiences (Dayter, 2014). Similar strategies are also observed by Ren and Guo (2020): comparing of oneself between one aspect and another and comparing of oneself between past and present. E.g. *"I must admit I am a slow learner, but this doesn't stop me from improving. My Spanish is now as fluent as a native speaker."*

- **Bragging as a complaint.** Speakers seem to complain about something whereas the bragging intention is embedded. However, it is not difficult for audiences to recognize their real purpose and the bragging act which is coached in the form of a complaint (Dayter, 2014; Ren and Guo, 2020). E.g. *"I am really tired of having the same free pie from the gym every day."*

Moreover, the success of self-presentation strategies requires "a delicate balance among self-enhancement, accuracy, and humility" (Schlenker and Leary, 1982) and is also impacted by the social context (Tice et al., 1995) or speaker identity (Paramita and Septianto, 2021).

2.3.2 Qualitative Studies on Bragging

Recent work shows that self-presentation is frequent, especially in digital communications (Dayter, 2014, 2018; Matley, 2018). Social media platforms tend to promote self-presentation (Chen et al., 2016) and allow users to craft an idealized self-image of themselves more conveniently (Chou and Edge, 2012; Michikyan et al., 2015; Halpern et al., 2017). Furthermore, self-promotion is acceptable and even desired in certain online contexts (Dayter, 2018). This is also amplified by social media platforms through the presence of likes or positive reactions to users' posts (Reinecke and Trepte, 2014) which often are used to quantify the impact on the platform (Lampos et al., 2014). Bragging in particular was found to be more frequent in social media than face-to-face interactions (Ren and Guo, 2020).

Bragging plays an important role in self-presentation and impression management. However, its pervasiveness challenges classic politeness theories, such as the modesty maxim (Leech, 2016) and the self-denigration maxim (Gu, 1990). Moreover, speakers' awareness of the socially risky nature of self-enhancement is manifested in the way they brag (Dayter, 2014). Thus, research in social psychology and linguistics has mostly focused on identifying the pragmatic strategies for bragging that mitigate face threats and their impact on likeability and perceived competence, which the speakers aim to increase with this self-presentation strategy.

The use of pragmatic strategies for bragging has been analyzed qualitatively by linguists and psychologists across languages such as Mandarin (Wu, 2011), English (Speer, 2012), Peninsular Spanish (Maíz-Arévalo, 2021), English and Russian (Dayter, 2021).

In social media, Dayter (2014) identified a series of overlapping strategies in a small ballet community on Twitter. The use of hyperlinks, images and hashtags, which is a unique pattern in social media, was also observed to serve for face mitigation.

Matley (2018) examined the pragmatic function of hashtags (e.g. #brag and #humblebrag) used by Instagram users. Users employ hashtags as illocutionary force indicating devices (IFIDs) to guide the overall meaning and interpretation of the utterance (Scott, 2015). In this way, users indicate their bragging straightforwardly by acknowledging their intentions, which makes them genuine and honest with the audience.

Tobback (2019b) focused on LinkedIn with special attention to the specific discursive strategies used in LinkedIn summaries (e.g. expressing in an objective way based on facts or figures). This is because LinkedIn users need to strengthen the elements of evidence to highlight their professional skillfulness while avoiding their self-praise being perceived negatively at the same time.

Rüdiger and Dayter (2020) identified three main brag types on a pick-up artist forum (i.e. a male community that shares techniques and scripts to quickly seduce women).

The frequency of each bragging strategy used by ordinary people (Ren and Guo, 2020) and celebrities (Guo and Ren, 2020) was examined on Chinese social media Weibo. According to their research, explicit bragging without modification is the most common type employed by ordinary Weibo users, followed by modified explicit bragging, while they infrequently use implicit bragging. In contrast, explicit bragging without modification is the least commonly used by celebrities and they predominantly choose modified explicit bragging and implicit bragging, where modified explicit bragging is used slightly more frequently. This is because celebrities need to seek a safe way to interact with and maintain their followers through positive self-presentation and pose fewer threats to followers' faces (Guo and Ren, 2020).

Also, some studies focus on bragging analysis from the perspective of the audience. Emotional influences of bragging recipients were investigated through an empirical study (Scopelliti et al. 2015). The results revealed that the degree to which recipients feel proud of and happy for speakers is overestimated while the degree to which recipients are annoyed at bragging is underestimated by speakers. Sezer et al. (2018) analyzed a special kind of bragging, humblebrag (i.e. bragging masked by complaint or humility which attempts to appear humble), on Twitter. Their studies showed that it is less effective and sincere than direct bragging and thus receives less liking. Later, Ren and Guo (2021) found that a popular online phenomenon (i.e. Versailles Literature) on Chinese social media Weibo is essentially a type of humblebrag by investigating its pragmatic strategies. Furthermore, audiences' reactions to literal and ironic bragging on Instagram were compared where the perceptions were higher for literal than for ironic ones in terms of sincerity, likability and modesty (Chalak, 2021).

Additionally, individual and cultural differences were examined in the related research. Women were found to be more likely than men to disclose more information about themselves offline (Maltz and Borker, 2018) and online (Barrett and Lally, 1999; Rosen et al., 2010; Rui and Stefanone, 2013) with more emotional exchanges (Gefen and Ridings, 2005). Rui and Stefanone (2013) compared the self-presentation and image management behaviors of American and Singaporean users on Facebook and found that Americans update their profiles more frequently while Singaporeans share significantly more photos. This might result from the differences between individualistic (e.g. Americans) and collectivistic (e.g. Singaporeans) culture: individualistic culture tends to focus on oneself while collectivistic culture aims to maintain the relationship and seek attention by sharing photos (Rui and Stefanone, 2013). Moreover, Moon et al. (2016) examined the relationship between narcissism (i.e. a personality trait reflecting a grandiose and inflated self-concept) and self-promoting behavior on Instagram. The results showed that Instagram users higher in narcissism tend to post more selfies and self-presented photos and update their profile pictures more often. Lastly, Tobback (2019a) observed less willingness for U.S. communication professionals to brag excessively than French ones on LinkedIn.

However, all these studies rely on manual analysis of small data sets (e.g. <300 posts). Also, individual differences (e.g. gender, age, education levels, personalities) in bragging behavior online have attracted much less research attention.

2.4 Detecting Speech Acts as Text Classification

Automatically identifying and classifying speech acts can be defined as a text classification task, which is a fundamental NLP task. It aims to automatically assign texts or documents to predefined labels (e.g. positive, negative in sentiment analysis) based on their content. Given a data set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ where x_i is the ith text in D and y_i is the corresponding target label, the goal is to make the prediction $\hat{y}_i = f(x_i)$ by training classification models.

Input The input texts can be social media posts, which tend to be presented in a short form with diverse topics and languages (Jin et al., 2013; Laksana and Purwarianti, 2014). Most of them contain grammatical mistakes, typos, emojis and informal terms such as abbreviations, slang, letter repetition and colloquial expressions (Panagiotou et al., 2016). Moreover, they may cover trending topics and terms. Overall, the usage of irregular expressions and up-to-date topics make classifying speech acts in social media a difficult task (Kolajo et al., 2020).

Models Traditional machine learning algorithms that are popular for text classification include logistic regression, support vector machines (SVMs; Cortes and Vapnik 1995) and neural networks. One of the main neural network models is recurrent neural networks (RNNs) where the previous outputs are fed as input to the current step. They use sequential data and represent contextual information through the surrounding context (Zuo et al., 2016).

Recently, pretrained transformer-based (Vaswani et al., 2017) language models have been applied to text classification tasks with promising results (Liu et al., 2019; Hoang et al., 2019; Shangipour ataei et al., 2020). A widely used one is BERT model (Bidirectional Encoder Representations from Transformers; Devlin et al. 2019), which is composed of multiple layers of transformer networks. Unlike recurrent networks, it uses positional embeddings to represent the position of the word in a sequence. It is trained on large corpora using masked language modeling (MLM), which randomly masks some of the tokens from the input aiming to predict the masked words based on the context only (Devlin et al., 2019). The MLM objective enables the model to learn deep bidirectional representations based on both left and right of the word. Following the success of BERT, other variants have been developed for different pretraining objectives (Lan et al., 2019; Sanh et al., 2019; Liu et al., 2019) and domain-specific tasks (Chalkidis et al., 2020; Lee et al., 2020; Nguyen et al., 2020).

Output A classification layer is added on top of the neural networks or transformer-based models for classification. It outputs the probabilities for each label (e.g. complaint, non-complaint) to obtain the predictive result.

2.5 Modeling Complaints in NLP

Previous work on automatic complaint identification has focused on using basic feature-based machine learning and neural network models (Coussement and Van den Poel, 2008; Jin et al., 2013; Preoţiuc-Pietro et al., 2019).

Coussement and Van den Poel (2008) improved the predictive performance in automatic complaint email classification by adding linguistic features (e.g. word count, number and negations) as an additional set of predictors.

Jin et al. (2013) proposed an enlargement method to improve the performance of predicting service failure complaints from a certain hotel website and found their method is helpful especially when the number of labeled samples is very small. They also compared two supervised machine learning algorithms, K-Nearest Neighbors (KNN) (Cover and Hart,

1967) and SVM (Cortes and Vapnik, 1995) in the complaint identification task, where SVM algorithm performed better than KNN.

Preoţiuc-Pietro et al. (2019) applied neural models and logistic regression with a wide variety of features, which include linguistic, sentiment, emotional and complaint-specific features (e.g. request, pronoun types), to identify complaints on Twitter. They also used a distant supervision method (i.e. pretrain on a larger "noisy" complaint data set) to further boost the performance.

Other studies have classified complaints into task-specific categories. Lailiyah et al. (2017) classified sentiment in Indonesian public complaints from Twitter and the government website. Also, Laksana and Purwarianti (2014) and Gunawan et al. (2018) divided Indonesian government complaints and customer complaints on Twitter into predefined categories respectively according to the responsible department. Other complaint-related categorizations are based on product hazards and risks from Amazon reviews (Bhat and Culotta, 2017) and escalation likelihood of customer complaints from a Chinese e-commerce company (Yang et al., 2019a).

In addition, He et al. (2014) proposed a measure model to calculate the influence of complaint text and theme on Chinese social media Weibo. Hu et al. (2019) investigated the content of complaints by comparing the topics discussed in positive reviews and negative reviews of hotels on Tripadvisor using a structural topic model. Moreover, Ekinci et al. (2016) explored the relationship between personality traits (e.g. extraversion, openness) and complaining intentions using logistic regression models.

Overall, most of the previous studies on modeling complaints in social media have focused on automatic complaint identification or task-specific complaint classification using feature-based machine learning models or task-specific neural models trained from scratch. Table 2.6 summarizes some limitations of previous work on automatic complaint detection. State-of-the-art pretrained neural language models and a generic complaint type classification task (e.g. severity levels) have not been explored before the work in this thesis. We compare the previous research and our work, highlighting their similarities and differences:

- Linguistic information such as emotion and topic was used in both some previous work and our work to boost performance.
- Most previous work limited complaints to a specific domain (e.g. hotel reviews, public complaints); complaints in our work span nine domains in social media.

Work	Limitations
Coussement and Van den Poel (2008)	The study focused on identifying complaints from emails rather than from social media platforms; The linguistic features used for performance boost were extracted from surface text information only; The data set was not publicly available.
Jin et al. (2013)	Complaints were collected exclusively from reviews related to one hotel; The use of traditional machine learning models (e.g. KNN, SVM) suggested a possibility for enhancing performance; The data set was not publicly available.
Laksana and Purwarianti (2014)	The categorizations of complaints were task-specific: classifying public complaints about a certain city government based on relevant government agencies; The data set was not publicly available.
Bhat and Culotta (2017)	The work restricted consumer complaints to one aspect: identifying a potential safety or health hazard of a product; They only manually labeled 448 reviews for validation and used unlabeled learning with domain adaption for training.
Lailiyah et al. (2017)	The work classified sentiment of public complaints instead of identifying complaints.
Gunawan et al. (2018)	The categorizations of complaints were task-specific: classifying customer complaints about a certain company into four divisions; The work used Naive Bayes classifier only without comparing other methods; The data set was not publicly available.
Yang et al. (2019a)	The use of traditional machine learning and neural models with feature engineering techniques suggested a possibility for enhancing performance; The work lacked detailed analysis of how models performed on predicting complaint escalation (e.g. error analysis, case study); The data set was not publicly available.
Preoţiuc-Pietro et al. (2019)	The use of traditional machine learning and neural models with feature engineering techniques suggested a possibility for enhancing performance; The work lacked detailed analysis of how models performed on identifying complaints (e.g. error analysis, case study).

Table 2.6: Limitations of previous work on automatic complaint classification.

- Resources in most previous work were not publicly available; we release the manually annotated data set as well as models.
- Previous work used traditional machine learning and neural network models with manually-engineered features; we use advanced transformer-based language models with injected features and MTL settings.

Later research (after our work) has also shown significant attention to enriching the study of automatic complaint classification to fill this gap (Singh et al., 2021b, 2022c; Fang et al., 2022; Bhatia et al., 2022). MTL settings were developed using neural models (Singh et al., 2022c) and pretrained models based on transformer networks (Singh et al., 2021b; Singh and Saha, 2021; Singh et al., 2022b) to boost the performance of complaint identification. They jointly model complaint identification (primary task) and one or more related tasks (auxiliary tasks) such as sentiment classification (Singh et al., 2021b, 2022c), emotion and sentiment detection (Singh and Saha, 2021; Singh et al., 2022a), emotion, sentiment and sarcasm detection (Singh et al., 2022b). Moreover, Singh et al. (2022a) used multi-task multi-modal (i.e. text and image) architectures for automatic complaint identification in Amazon reviews; while Bhatia et al. (2022) applied MTL frameworks to classify complaints and their severity levels in financial domain, which was assisted by emotion and sentiment classification task.

Semi-supervised approaches were adapted for complaint detection on Twitter (Gautam et al., 2020; Singh et al., 2021a). Complaints were also modeled and analyzed in different languages (Singh et al., 2020; Nguyen et al., 2021; Fang et al., 2022; Ito et al., 2022). Singh et al. (2020) and Nguyen et al. (2021) evaluated a wide range of models including semi-supervised models, traditional machine learning models, deep learning models and pretrained models on the task of complaint identification in Hindi and Vietnamese respectively. Moreover, Fang et al. (2022) analyzed the emotional intensity of complaints on Chinese social media Weibo; while Ito et al. (2022) classified the target scope of complaints (i.e. self, individual, group, environment) on Twitter.

Overall, our work has made substantial contributions to the field of complaint identification and classification. It has not only demonstrated superior performance in accurately and efficiently identifying complaints but also sparked further research in the area of generic complaint classification. It allows for a more comprehensive understanding of complaints, which potentially leads to more informed decision-making and enhanced customer satisfaction.

2.6 Modeling Self-Disclosure in NLP

To the best of our knowledge, there is no previous work on modeling bragging in NLP. Closely related to bragging, self-disclosure is a communication process used for revealing emotions and personal information about oneself to others (Bak et al., 2012). It is usually employed to pursue social rewards such as maintaining or improving relationships and increasing social support (Duck, 1998; Bak et al., 2014).

Bak et al. (2012) investigated the relationship between self-disclosure and relationship strength (e.g. strong and weak relationships) based on the duration and frequency of interactions on Twitter using text mining techniques. The results showed that Twitter users tend to disclose more to their close friends (i.e. high relationship strength) while they share more positive sentiment with weak relationships.

Jaidka et al. (2018) compared self-disclosure behavior on Facebook and Twitter from the perspective of users' demographic and psychological traits. Their results revealed that users prefer to self-disclose more on Facebook than on Twitter, especially the information about their family, personal concerns and emotions, while users are more likely to share their ambitions and goals on Twitter.

Umar et al. (2019) used text mining methods to detect and analyze self-disclosure in newspaper comment forums. They also examined the effects of anonymity and topic of discussion on self-disclosure behavior and found that anonymous users are more likely to disclose about themselves than identifiable users.

Umar et al. (2021) focused on studying self-disclosure on Twitter during the coronavirus pandemic. They used an unsupervised approach for self-disclosure detection and compared them with self-disclosure during Hurricane Harvey (in 2017).

Previous work also detected different levels of self-disclosure (e.g. high, low and no self-disclosure) from Twitter conversations (Bak et al., 2014) and health-related posts online (Balani and De Choudhury, 2015; Valizadeh et al., 2021).

More recently, Wang et al. (2021) identified self-promotion (i.e. presenting oneself as competent) by U.S. Congress members on Twitter and examined gender differences in self-promotion using machine learning models. Moreover, bragging in some cases also involves possessions, which have been mined from texts in the past research (Chinnappa and Blanco, 2018).

Work	**Limitations**
Bak et al. (2012, 2014)	The work lacked certain clarifications concerning data collection, such as whether the dyads of Twitter users engaged in conversations were friends or not; The work lacked cases of model behaviors (e.g. error analysis); The data sets were not publicly available.
Balani and De Choudhury (2015); Valizadeh et al. (2021); Umar et al. (2021)	The study focused exclusively on health-related self-disclosure or self-disclosure during the COVID-19 pandemic. Data sets were not publicly available (except Valizadeh et al. 2021)
Jaidka et al. (2018)	The use of traditional machine learning models (e.g. LR, SVM) suggested a possibility for enhancing performance; Demographic traits explored in this study included only age and gender; The data set was not publicly available.
Umar et al. (2019)	The study focused exclusively on detecting self-disclosure from user comments on news articles; The work lacked cases of model behaviors (e.g. error analysis); The data set was not publicly available.
Wang et al. (2021)	The work focused exclusively on self-promotion of Congresspeople in the U.S.; The work only explored the relationship between self-promotion behavior and gender.

Table 2.7: Limitations of previous work on modeling self-disclosure or self-promotion.

Table 2.7 summarizes some limitations of previous work on modeling self-disclosure or self-promotion. We compare the previous work and our work, highlighting their similarities and differences:

- Both previous studies and our work are based on linguistic or social psychological theories (e.g. conceptualization). Also, findings in previous work and our work have been validated through linguistic or social psychology studies.

- Previous research concentrated on modeling self-disclosure; our work focuses on investigating bragging, one of the forms of self-disclosure.

- Previous research explored only one or two demographic traits (e.g. age, gender) in self-disclosure/self-promotion studies; our work examines the relationship between bragging behavior online and a series of user traits including age, gender, education and income.

Thus, we view our work as the pioneering effort to apply the concept of bragging to computational linguistics using pragmatic theory. We introduce methods for computational analysis of bragging in social media and demonstrate encouraging findings. Our findings support previous sociolinguistics studies with more robust results from a large-scale data set and show the effectiveness of computationally analyzing bragging behavior online.

2.7 Summary

This chapter presented a background on speech acts focusing on complaining and bragging. We provided the definitions as well as their types and pragmatic strategies. Then, we introduced the previous work on analyzing complaining and bragging (or self-presentation) in linguistic, psychology and NLP research. Based on this, we noticed that (1) previous studies on modeling complaining in NLP have focused on complaint identification and task-specific classification using traditional machining learning and neural models with feature engineering; (2) bragging has only been manually analyzed on a small scale in linguistic and psychology studies. There is a need for studying bragging at scale in computational linguistics, introducing novel complaint classification tasks (i.e. complaint severity classification) and developing more advanced transformer-based models for these tasks.

Chapter 3

Modeling the Severity Level of Complaints

In Section 2.2, complaining has been defined as a speech act that usually conveys negative emotions triggered by a discrepancy between reality and expectations towards an entity or event (Olshtain and Weinbach, 1987). Complaints play an important role in human communication for expressing dissatisfaction. In pragmatics, complaints have been classified into various levels of severity according to the amount of face-threat that the complainer is willing to undertake and their purpose (e.g. express dissatisfaction, find solutions) (Olshtain and Weinbach, 1987; Trosborg, 2011; Kakolaki and Shahrokhi, 2016) (see Section 2.2.1).

Recent work on modeling complaints in NLP has focused on distinguishing complaints from non-complaints in social media (Jin et al., 2013; Preoţiuc-Pietro et al., 2019). However, there is no previous study dividing them into more fine-grained generic categories (see Section 2.5). Table 3.1 shows examples of social media posts expressing complaints grouped into four severity classes according to Trosborg (2011): (a) no explicit reproach; (b) disapproval; (c) accusation; and (d) blame.

Identifying and analyzing the severity of complaints is important for: (a) improving customer service by recognizing the level of dissatisfaction and understanding complainers' needs (Au et al., 2009; Vásquez, 2011); (b) linguists to study the speech act of complaints in different levels of granularity on large scale (Boxer, 1993a; Noisiri, 2002); and (c) developing downstream NLP applications such as automatic complaint response generation (Xu et al., 2017) or voting stance prediction (Tsakalidis et al., 2018).

Label	**Example**
No Explicit Reproach	*Are you following me? I seem unable to send you a dm.*
Disapproval	*So far, the mac graphics drivers have been another disappointing update (for both my quadro 4000 & gtx - 285),*
Accusation	*Can u stop adding the UK keyboard layout to my Italian keyboard at every update? ktnxby*
Blame	*Thanks to <USER>'s incompetence i now can't work till October 4th, when the ati card arrives.*

Table 3.1: Examples of complaint severity levels.

The main contributions of this chapter are as follows: (1) grounded in the linguistic theory of pragmatics (Trosborg, 2011), we enrich a publicly available data set (Preoţiuc-Pietro et al., 2019) with four complaint severity levels; (2) we create a new classification task for identifying different severity levels of complaints; (3) we evaluate transformer-based classification models (Vaswani et al., 2017) combined with linguistic information on complaint severity level classification.

The work presented in this chapter has been published at the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021) (Jin and Aletras, 2021).

3.1 Task & Data

We define complaint severity prediction as a multi-class classification task. Given a text snippet T, defined as a sequence of tokens $T = \{t_1, ..., t_n\}$, the aim is to classify T as one of the four predefined severity labels.

We use an existing complaints data set developed by Preoţiuc-Pietro et al. (2019), which consists of 1,232 complaints (35.7%) and 2,217 non-complaints (64.3%) in English. We opt to use this data set because it is publicly available with annotated complaints collected from Twitter in 9 general domains (i.e. Food, Apparel, Retail, Cars, Service, Software, Transport, Electronics and Other).

3.1.1 Complaint Severity Categories

For complaint severity annotation, we adopt the four categories defined by Trosborg (2011) because it is widely accepted and used in other linguistic studies (Noisiri, 2002; Al-Shboul, 2021) (see examples in Table 3.1):

- **No explicit reproach:** there is no explicit mention of the cause and the complaint is not offensive;
- **Disapproval:** express explicit negative emotions such as dissatisfaction, annoyance, dislike and disapproval;
- **Accusation:** asserts that someone did something reprehensible;
- **Blame:** assumes the complainee is responsible for the undesirable result.

Note that the severity levels categorize complaints by type instead of intensity. Classes are disjoint according to the definitions by Trosborg (2011). More specifically, *"No explicit reproach"* is a suggestive strategy, where the complainee is usually not mentioned in the statement. *"Disapproval"* expresses negative sentiment or an unsatisfying state only. The statement may imply that the complainee is responsible for the dissatisfying result but avoid mentioning it, which is the key component of identifying "Disapproval" and "Accusation"/"Blame". The main difference between *"Accusation"* and *"Blame"* is that the complainer presupposes the complainee is guilty of the offense in a "Blame" complaint.

3.1.2 Complaint Severity Annotation

Following the definitions above, each tweet was labeled by three annotators independently. In case of ties, the final decision was made by the authors through consensus. We recruited 35 native English-speaking annotators from the volunteer list of our institution (a volunteer list with all students in the computer science department). All volunteers were unpaid and received no rewards due to the limited budget. Annotators were provided with guidelines including an introduction of the task and definitions and examples of each category (see

Labels	**Amount**	**Percentage (%)**
No Explicit Reproach	436	35.4
Disapproval	376	30.5
Accusation	224	18.2
Blame	196	15.9
Total	1,232	100

Table 3.2: Statistics of complaint severity level data set.

Appendix A). The inter-annotator agreement between three annotations is (a) percentage agreement: 60.28; (b) Krippendorff's Alpha (Krippendorff, 2011): 0.634, which belongs to *substantial* agreement (Artstein and Poesio, 2008).

Table 3.2 shows the distribution of tweets across four classes: 436 tweets belong to "No Explicit Reproach" (35.4%), 376 belong to "Disapproval" (30.5%), 224 belong to "Accusation" (18.2%) and 196 belong to "Blame" (15.9%). The class distributions over 5 domains (Car, Retail, Service, Software, Transport) are similar to the overall distribution while 4 domains (Food, Apparel, Electronics, Other) differ from Table 3.2. In domains with different distributions, differences appear especially in "No Explicit Reproach" and "Accusation", which might result from domain-specific complaint requests.

3.1.3 Text Processing

Text is processed by lower-casing and replacing all mentions of usernames and URLs with placeholder tokens <USER> and <URL> respectively. A Twitter-aware tokenizer, DLATK (Schwartz et al., 2017), is used for text tokenization to handle emoticons and hashtags in social media texts.

3.2 Methodology

We evaluate a pretrained transformer-based model and its combination with linguistic information (i.e. emotion and topic information) on the severity complaint classification task.

3.2.1 RoBERTa

Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2019) is a pre-trained language model based on the Transformer architecture (Vaswani et al., 2017). It makes use of multiple multi-head attention layers to learn context information from both the left and the right sides of tokens. It is trained on masked language modeling by randomly masking some of the tokens from the input aiming to predict them based on the context only. RoBERTa (Liu et al., 2019) is an extension of BERT trained on more data with different hyperparameters and has achieved better performance in social media analysis tasks (Maronikolakis et al., 2020). We fine-tune RoBERTa[1] on complaint severity classification by adding an output dense layer with a softmax activation function.

3.2.2 RoBERTa with Linguistic Information

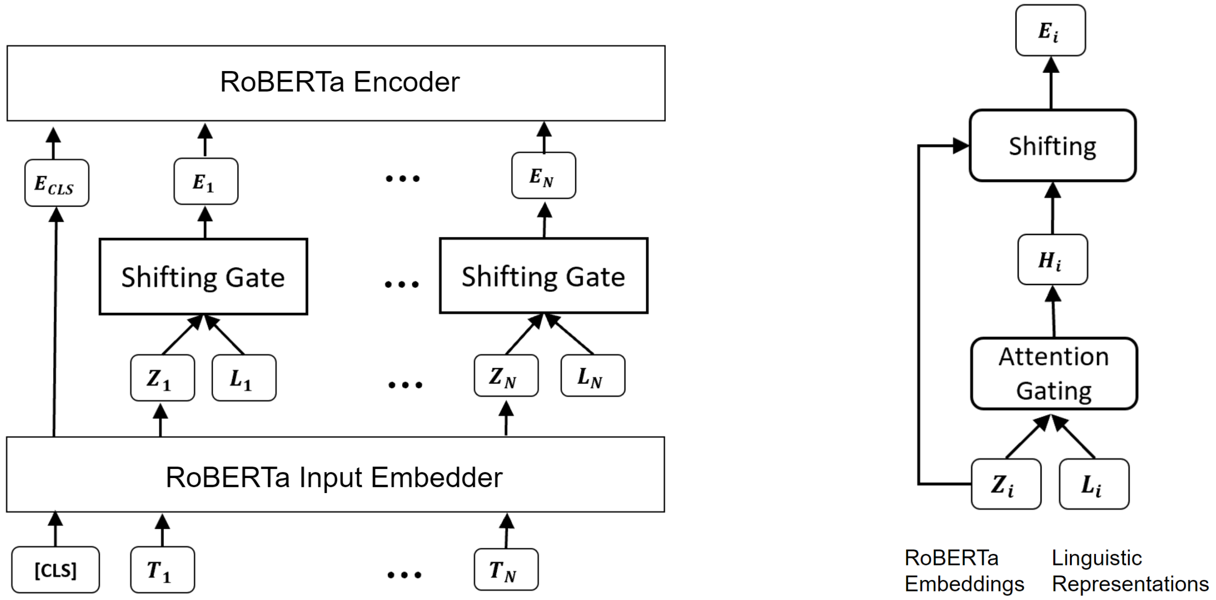
Rahman et al. (2020) proposed a method that injects multimodal information such as image and speech into the text representations of BERT model. It combines word embeddings and embeddings from other modalities (e.g. image, audio) and then feds the combination to a BERT encoder. We adapt it by replacing (1) the underlying BERT model with RoBERTa model; and (2) the multimodal information with linguistic information. The architecture of RoBERTa with linguistic information network is shown in Figure 3.1a.

We first use a fully connected layer to project the linguistic representations into vectors with comparable size to RoBERTa's embeddings. Then we concatenate word representations Z_i obtained from RoBERTa and linguistic information representations L_i using a shifting gate (Wang et al., 2019) called Multimodal Adaption Gate (MAG; see Figure 3.1b), where an attention gating mechanism is applied to control the influence of each representation. The following equation shows the operation process to generate the gating vector g_i^l in Attention Gating:

$$g_i^l = R(W_{gl}[Z_i; L_i]) \tag{3.1}$$

where $[Z_i; L_i]$ is the combined embeddings, W_{gl} is a learnable weight matrix and $R(x)$ is a non-linear activation function (ReLU). Then a non-verbal displacement vector H_i is computed

[1] We only report the results of RoBERTa because it achieves better performance compared to BERT over all evaluation methods in our experiments.



(a) The overview of RoBERTa with linguistic information network.

(b) The architecture of a shifting gate.

Figure 3.1: The architecture of RoBERTa with linguistic information network.

using the following equation:

$$H_i = g_i^l \cdot (W_l L_i) \tag{3.2}$$

where W_l is a learnable weight matrix for linguistic information. The final combined vector E_i is calculated by adding Z_i and αH_i together:

$$E_i = Z_i + \alpha H_i \tag{3.3}$$

$$\alpha = min(\frac{\|Z_i\|_2}{\|H_i\|_2}\beta, 1) \tag{3.4}$$

where β is a hyperparameter and $\|Z_i\|_2$ and $\|H_i\|_2$ refer to the L_2 norm of Z_i and H_i respectively. Finally, we apply layer normalization and dropout after the shifting gate and pass the output to a RoBERTa encoder. We add an output layer to RoBERTa for classification which is similar to the RoBERTa model. We use RoBERTa with three types of linguistic features (i.e. emotion, topic and their combination) to explore whether these can provide extra information that word embeddings are unable to extract.

RoBERTa$_{Emo}$ We first use emotional information obtained by using a pretrained emotional classifier by Volkova and Bachrach (2016). This is a 9-dimensional vector representing scores of sentiment (positive, negative and neutral) and six basic emotions of Ekman (1992) (anger, disgust, fear, joy, sadness and surprise).

RoBERTa$_{Top}$ We also use topical information from a 200-dimensional vector representing the distribution of the fraction of tokens in each tweet belonging to a topic cluster (Preoţiuc-Pietro et al., 2015b).

RoBERTa$_{Emo+Top}$ We finally experiment with injecting both emotional and topical information to RoBERTa.

3.3 Experimental Setup

3.3.1 Baselines

Majority Class We use Majority Class as the first baseline, where we calculate scores by labeling all the tweets with the majority class.

LR-BOW We use a linear baseline, Logistic Regression with standard bag-of-words (LR-BOW) and L2 regularization.

BiGRU-Att We also use a neural baseline trained from scratch, a bidirectional Gated Recurrent Unit (GRU) network (Cho et al., 2014) with a self-attention mechanism (BiGRU-Att; Tian et al. 2018). Given a Twitter post T, a token t_i is mapped to a GloVe embedding (Pennington et al., 2014). We then apply dropout to the output of GloVe embedding layer and pass it to a bidirectional GRU with self-attention layer. Finally, the contextualized token representations are passed to an output layer using a softmax activation function for multi-class classification.

3.3.2 Hyperparameters

The **BiGRU-Att** model uses 200-dimensional GloVe embeddings (Pennington et al., 2014) pre-trained on Twitter data. Its hidden size is $h = 128$, $h \in \{64, 128, 256, 512\}$ with dropout $d = .2$, $d \in \{.2, .5\}$. We use Adam optimizer (Kingma and Ba, 2014) with learning rate $l =$ 1e-3, $l \in \{$1e-3, 5e-3, 1e-2$\}$. For **RoBERTa**, we use the base uncased model and fine-tune it with learning rate $l =$ 5e-6, $l \in \{$1e-4, 1e-5, 5e-6, 1e-6$\}$. The maximum sequence length is set to 50 covering 95% of tweets in the training set. For **RoBERTa with linguistic features**, we project the linguistic features (emotions and topics) to vectors of size $l = 200$, $l \in \{200, 300, 400, 768\}$. We also use dropout $d = .5$, $d \in \{.2, .5\}$. For the shifting gate MAG, we use the default parameters from Rahman et al. (2020). For all models, we use a categorical-cross entropy loss following a similar approach to Sun et al. (2019) which has achieved the best results on fine-grained sentiment analysis (i.e. similar to the ordinal scale of complaints severity).

3.3.3 Training and Evaluation

We run all models using a nested 10-fold cross validation approach, which consists of 2 nested loops as in Preoţiuc-Pietro et al. (2019). In the outer loop, 9 folds are used for training and one for testing; while in the inner loop, a 3-fold cross validation method is applied to the data from the nine folds (in the outer loop), where 2 folds are used for training and one for validation. During training, we choose the model with the smallest validation loss over 30 epochs. We measure predictive performance using the mean Accuracy, Precision, Recall and macro F1 over 10 folds. We also report the standard deviations.

3.4 Results

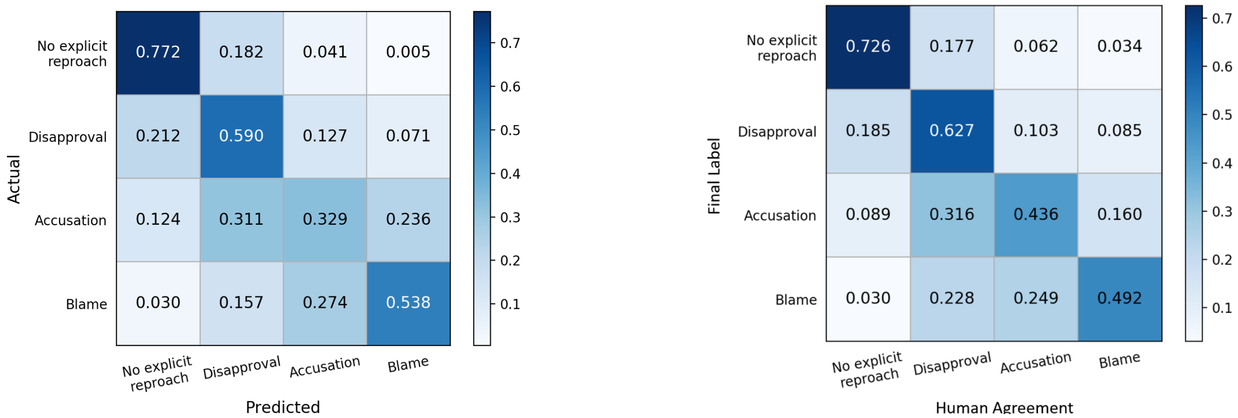
Table 3.3 shows the performance of all models including baselines, RoBERTa model and RoBERTa combined with linguistic information on complaint severity level prediction.

Model	**Acc**	**P**	**R**	**F1**
Majority Class	35.2	8.8	25.0	13.0
LR-BOW	46.7 ± 2.8	44.3 ± 2.8	43.6 ± 2.9	43.5 ± 2.7
BiGRU-Att	46.1 ± 2.7	43.6 ± 3.1	42.7 ± 2.4	41.8 ± 2.4
RoBERTa	58.7 ± 2.8	55.8 ± 5.2	55.4 ± 3.4	54.7 ± 4.0
RoBERTa$_{Emo}$	**59.8** ± 3.3	**56.6** ± 3.7	55.7 ± 3.9	**55.7**† ± 3.8
RoBERTa$_{Top}$	59.0 ± 3.4	55.9 ± 4.0	55.6 ± 3.2	55.2 ± 3.8
RoBERTa$_{Emo+Top}$	59.4 ± 2.9	56.5 ± 2.8	**56.2** ± 2.7	55.5 ± 2.5

Table 3.3: Accuracy (Acc), Precision (P), Recall (R) and macro F1-Score (F1) for complaint severity level prediction (± std. dev.). Best results are in bold. † indicates statistically significant improvement over RoBERTa (t-test, $p < 0.05$).

Overall, RoBERTa with linguistic features achieves the best results. **RoBERTa**$_{Emo}$ outperforms all other models and reaches macro F1 up to 55.7. This confirms that injecting extra emotional information helps improve the performance of complaint severity level prediction. This is also in line with Trosborg (2011) who states that the expression of complaints is relevant to different emotional states. The results of **RoBERTa**$_{Top}$ and **RoBERTa**$_{Emo+Top}$ are comparable with 55.2 and 55.5 macro F1 respectively. **RoBERTa** performs competitively but worse than the RoBERTa with linguistic features. We also notice that **BiGRU-Att** does not perform well in our task (41.8 macro F1), which may result from the fact that it does not take into account word order during training.

Figure 3.2a presents the confusion matrix of our best model (i.e. **RoBERTa**$_{Emo}$). The confusion matrix is normalized over the actual values (rows). The **"No Explicit Reproach"** category has the highest percentage (77.2%) of correctly classified data points by the model, followed by the label **"Disapproval"** with 59.0%. These are also the two most frequent classes in the data set. On the other hand, results of **"Accusation"** are the lowest (32.9%) which is confused with adjacent categories ("Disapproval" and "Blame"). Furthermore, the differences between misclassifications and correct classification are relatively large for **"Blame"**. We speculate that this is because of the unique linguistic characteristic of the "Blame" category which gives emphasis on someone's responsibility. Finally, a category is more likely, in general, to be misclassified to its adjacent severity categories. For example, when predicting **"Disapproval"**, the number of model misclassifications as "No Explicit



(a) Confusion matrix of the best performing model (RoBERTa$_{Emo}$).

(b) Confusion matrix of human agreement.

Figure 3.2: Confusion matrices from modeling the severity level of complaints.

Reproach" and "Accusation" is larger than "Blame". This hints that tweets belonging to neighboring levels share more semantic, syntactic and stylistic similarities.

We also compare the performance of RoBERTa$_{Emo}$ with the human agreement for each class (see Figure 3.2b). In general, the results of the model correlate to human agreement. In other words, the model and humans agree on the categories they confuse. For instance, it is easy for both of them to confuse **"Accusation"** with **"Disapproval"** (32.9% vs. 31.1% for the model and 43.6% vs. 31.6% for humans). However, we observe that annotators are better at distinguishing high severity complaints from **"No Explicit Reproach"**, where 21.2% "Disapproval" and 12.4% "Accusation" are wrongly classified as "No Explicit Reproach" by the model while the corresponding values are 18.5% and 8.9% by humans respectively. We argue that this is because annotators are able to identify the subtle language (more details will be discussed in Section 3.5). Also, we notice that the model achieves better performance when predicting **"Blame"**, indicating a better capability on capturing the main characteristics of this class compared to humans.

3.5 Discussion

We perform an error analysis to shed light on the limitations of our best performing model (**RoBERTa**$_{Emo}$) on complaint severity level classification.

Firstly, we observe that most errors happen when tweets belonging to **"Accusation"** share more similarities with **"Disapproval"** and **"Blame"**. The following two tweets are typical examples of "Accusation" being misclassified as "Disapproval" and "Blame" respectively:

> T1: *"<USER>, thank you! Clear guidelines here, but* **not at all** *what your advisor on the phone stated!"*

> T2: *"The new* <**USER**> *stinks ...10mins to* **take my order** *and another 15 to get it. And* **stop asking** *my name like we're friends <URL>"*

This is because some tweets belonging to **"Accusation"** also contain negation (e.g. *not at all*) or negative terms (e.g. *disappointed*), which appear frequently in **"Disapproval"**. Also, consistent with the definition by Trosborg (2011) (directly or indirectly accuses someone for causing the problem), tweets belonging to **"Accusation"** may involve doing something and contain terms like "<USER>" or "you", which is similar to complaints labeled as **"Blame"** such as:

> T3: *"Thanks* <**USER**> *for* **selling** *expired beer #fail* <**USER**> *<URL>"*

Secondly, the model struggles with complaints expressed in more subtle ways. In the following two examples, tweets belonging to **"Disapproval"** and **"Accusation"** are misclassified as **"No Explicit Reproach"** respectively:

> T4: *"Think someone at <USER> had been drinking the stuff before they put the label on"*

> T5: *"Just opened a fresh bud light that was filled with water. Please explain <USER>."*

Such complaints do not contain terms that are typical of any specific complaint severity category (e.g. negation and negative terms in **"Disapproval"**, person pronouns and terms describing undesirable results in **"Blame"**) thus predicting them correctly needs more contextual understanding.

Finally, compared to other categories, the model is more likely to confuse tweets belonging to **“No Explicit Reproach”** and **“Disapproval”**. This happens because some tweets express weak dissatisfaction, which is difficult to identify. The following tweet is misclassified as **“No Explicit Reproach”**:

> T6: *“Dearest <USER>: there really needs to be an easier method to report names that are inappropriate <URL>”*

The model might need to learn more contextual information about such tweets instead of capturing certain relevant terms. Also, these two labels contain more similar terms such as *dm*, *please help*, *can’t work* and interrogative tone. Examples of a **“No Explicit Reproach”** and **“Disapproval”** complaint are the following (where similarities are in bold):

> T7: *“Hey guys, I love this product featured on <USER> today* **but** *don’t see a price?* **Help** *a girl out***?** *<URL>”*

> T8: *“So it’s going to cost $7000 to fix the exhaust on my <USER> 2009 jetta, and* **only** *$300 is covered under warranty.* **Help** *<USER>***?***”*

3.6 Summary

In this chapter, we presented the first study on the severity level of complaints in computational linguistics. We developed a publicly available data set of tweets labeled with four categories based on the theory of pragmatics. We proposed an approach that allows the injection of linguistic features into transformer-based networks. Then, we modeled complaint severity level prediction as a new multi-class classification task and conducted experiments using the proposed models with different linguistic features. The results showed that adding emotional and topical information is beneficial to predicting the severity levels of complaints. Finally, through error analysis, we found that models struggled with subtle expressions and texts belonging to neighboring levels.

The method to inject linguistic information into transformer-based models is also employed in Chapters 4 and 5 for different tasks.

Chapter 4

Complaint Identification with Transformer Networks

We have previously introduced the task of complaint severity level classification in Chapter 3. We have also demonstrated in Section 2.5 that previous work on automatically identifying complaints in social media has focused on using feature-based and task-specific neural network models only. Adapting state-of-the-art pre-trained neural language models (Devlin et al., 2019; Liu et al., 2019) and incorporating other linguistic information (e.g. topic, emotion) for complaint prediction or jointly modeling the complaint identification task and other related tasks (i.e. MTL) have yet to be explored. Furthermore, as stated in Chapter 3, improving complaint identification accuracy is vital for customer service (Au et al., 2009; Vásquez, 2011), linguistic research (Boxer, 1993a; Noisiri, 2002) and downstream NLP applications (Xu et al., 2017; Tsakalidis et al., 2018).

In this chapter, we evaluate a series of neural models underpinned by transformer networks which we subsequently combine with linguistic information. We also model complaints with the help of severity level information in MTL settings.

The main contributions of this chapter are as follows: (1) we evaluate transformer-based classification models with the injection of linguistic information and a distant supervision method (first train models on a noisy but larger complaint data set to boost the predictive performance; Preoţiuc-Pietro et al. 2019) on complaint identification; (2) we achieve new state-of-the-art results on complaint identification in a multi-task setting; (3) we present a thorough analysis of limitations of transformers in predicting accurately whether a given text is a complaint or not.

The work presented in this chapter has been published at the Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020) (Jin and Aletras, 2020) and the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021) (Jin and Aletras, 2021).

4.1 Task & Data

Given a text snippet (i.e. tweet), we aim to classify it as a complaint or non-complaint. For that purpose, we use the data set by Preoţiuc-Pietro et al. (2019) described in Section 3.1. It contains tweets written in English that were manually annotated as complaints or not. It includes 1,232 complaints (62.4%) and 739 non-complaints (37.6%) over 9 domains (i.e. Food, Apparel, Retail, Cars, Services, Software, Transport, Electronics, Other). The statistics for each domain are shown in Table 4.4. To maintain a balanced data set, 1,478 non-complaint tweets were additionally sampled from replies and random tweets. In total, the data set includes 1,232 complaints (64.3%) and 2,217 non-complaints (35.7%). We opt to use this data set because (1) it is publicly available; and (2) it allows a direct comparison with existing methods.

We also use an extra complaint data set for distant supervision collected by Preoţiuc-Pietro et al. (2019). This larger but "noisy" data source contains 18,218 complaint tweets collected by querying Twitter API with certain complaint-related hashtags (e.g. #badbusiness, #badcustomerservice, etc.) and the same amount of non-complaint tweets that were sampled randomly. We use this distantly supervised data to first adapt all models on the complaint classification task. Then we fine-tune them using the smaller original complaint data set.

Domain	Complaints	Non-complaints
Food	95 (7.7%)	35 (4.7%)
Apparel	141 (11.5%)	117 (15.8%)
Retail	124 (10.1%)	75 (10.2%)
Cars	67 (5.4%)	25 (3.4%)
Services	207 (16.8%)	130 (17.6%)
Software	189 (15.3%)	103 (13.9%)
Transport	139 (11.3%)	109 (14.7%)
Electronics	174 (14.1%)	112 (15.2%)
Other	96 (7.8%)	33 (4.5%)
Total	1232	739

Table 4.1: Data set statistics across 9 domains.

4.2 Methodology

4.2.1 Single-Task Learning Settings

We first evaluate a series of transformer-based models as well as their combination with external linguistic information and distant supervision methods for complaint identification in single-task learning (STL) settings.

BERT, ALBERT and RoBERTa Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2019) learns language representations by jointly conditioning on both left and right contexts using transformers. It is trained on masked language modeling where some of the tokens are randomly masked with the aim to predict them using only the context.

We further experiment with ALBERT (Lan et al., 2019) and RoBERTa (Liu et al., 2019). ALBERT uses two parameter-reduction methods to address memory limitations and long training time of BERT: (a) factorized embedding parameterization; (b) cross-layer parameter sharing. RoBERTa is an extension of BERT trained on more data with a larger batch size using dynamic masking (i.e. changeable masked tokens of each sequence during training epochs). We adapt BERT, ALBERT and RoBERTa by adding a linear layer with a sigmoid activation and then fine-tune it on the complaint classification data.

XLNet XLNet (Yang et al., 2019b) uses a similar architecture to BERT to learn bidirectional contextual information. Instead of masked tokens used in BERT, XLNet maximizes the expected log-likelihood of all possible factorization orders. We adapt and fine-tune the XLNet model for complaint prediction similar to BERT.

BERT with linguistic features To combine our model with external linguistic information, we adapt the Multimodal BERT model structure (Rahman et al., 2020) that has been introduced for multimodal modeling (e.g. text, image, speech). Instead of cross-modal interactions, we inject extra linguistic information as alternative views of the data into the pre-trained BERT model as described in Section 3.2.2. We use (a) $\mathbf{BERT}_{Emo}$, a 9-dimensional vector obtained by quantifying sentiment (positive, negative and neutral) and six basic emotions of Ekman (1992) (anger, disgust, fear, joy, sadness and surprise) for each tweet using a predictive model by Volkova and Bachrach (2016); (b) $\mathbf{BERT}_{Top}$, a 200-dimensional vector representing word frequencies in word clusters designed to identify semantic themes in tweets by Preoţiuc-Pietro et al. (2015a); (c) $\mathbf{BERT}_{Emo+Top}$, both emotional and topical features. We first project the linguistic information into vectors with a similar size to BERT CLS embeddings. Then we concatenate word representations obtained from BERT and the linguistic features to generate combined representations. During concatenation, a shifting gate (Wang et al., 2019) called Multimodal Adaption Gate (MAG) is applied to control the importance of each representation. Finally, the output of the shifting gate is fed to a BERT encoder for fine-tuning. The rest of the architecture is the same as BERT.

4.2.2 Multi-Task Learning Settings

We further experiment with MTL (Caruana, 1997) for using severity categories (see Section 3.1.1) to improve binary complaint prediction (i.e. complaint or non-complaint). MTL enables two or more tasks to be learned jointly by sharing information and parameters of a model. We explore whether or not the severity level of a complaint helps in complaint identification[1]. We use the same data set where each tweet is annotated as a complaint or not and severity levels.[2]

We first adapt three multi-task learning models based on bidirectional recurrent neural networks proposed by Rajamanickam et al. (2020) for jointly modeling abusive language detection and emotion detection. We also adapt the best performing model (i.e. RoBERTa$_{Emo}$) in the complaint severity classification task (see Section 3.4) in a multi-task setting using two variants. We use the severity complaint prediction as an auxiliary task and the binary complaint prediction as the main task to train different MTL models. All models are trained on the two tasks and updated at the same time with a joint loss:

$$L = (1 - \alpha)L_{com} + \alpha L_{sev} \tag{4.1}$$

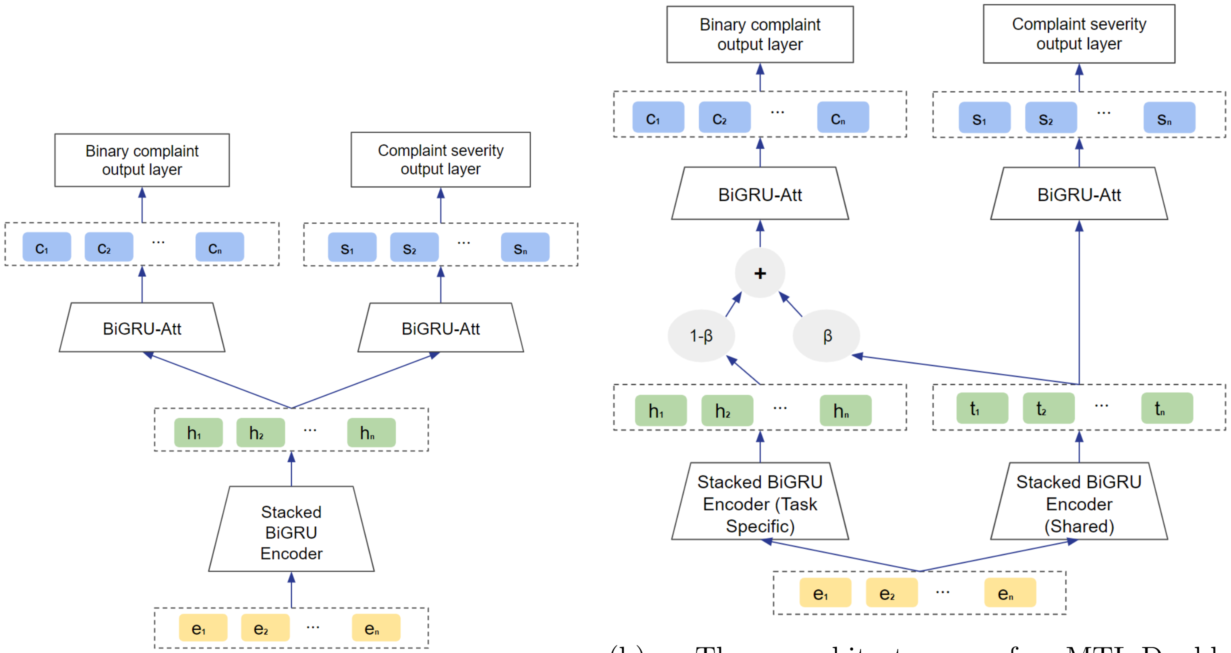
where L_{com} and L_{sev} are the losses of complaint identification and severity level classification tasks respectively. α is a parameter to control the importance of each loss.

MTL-Hard Sharing We adapt the MTL-Hard Sharing model of Rajamanickam et al. (2020), where a single encoder is shared and updated by both tasks. We first pass GloVe embedding representations to a shared stacked BiGRU encoder. Then the output of the shared encoder is fed to two different BiGRU-Att models specific to each task (complaint detection and severity level identification) separately. Finally, we add an output layer with a sigmoid and a softmax activation function for binary and multi-class prediction respectively (see Figure 4.1a).

MTL-Double Encoder Instead of sharing a single encoder, the MTL-Double Encoder model (Rajamanickam et al., 2020) uses two stacked BiGRU encoders, where one is task-specific (complaint identification only) and the other one is shared by both tasks. We pass the output of the shared encoder to a BiGRU-Att model for severity level prediction. We

[1] Initially, we planned to use MTL settings to jointly model severity levels of complaints (main task) and emotion or sentiment (auxiliary task) but this approach did not outperform transformer based models with directly injected emotion features. We guess assigning an emotion or sentiment label predicted by a pre-trained model can introduce noise, potentially contributing to the performance drop. We then attempted to replace the emotion or sentiment classes with binary complaint classes, but this modification did not lead to improved results. Finally, we made a significant change by swapping the main task (complaint identification) and the auxiliary task (complaint severity level classification) in the MTL settings.

[2] For a tweet that is a non-complaint, we assign an extra class for severity (i.e. "No Complaint Severity").


Binary complaint output layer
Complaint severity output layer
c₁
c₂
...
cₙ
s₁
s₂
...
sₙ
BiGRU-Att
BiGRU-Att
h₁
h₂
...
hₙ
Stacked BiGRU Encoder
e₁
e₂
...
eₙ
Binary complaint output layer
Complaint severity output layer
c₁
c₂
...
cₙ
s₁
s₂
...
sₙ
BiGRU-Att
BiGRU-Att
+
1-β
β
h₁
h₂
...
hₙ
t₁
t₂
...
tₙ
Stacked BiGRU Encoder (Task Specific)
Stacked BiGRU Encoder (Shared)
e₁
e₂
...
eₙ

(a) The architecture of MTL-Hard Sharing.

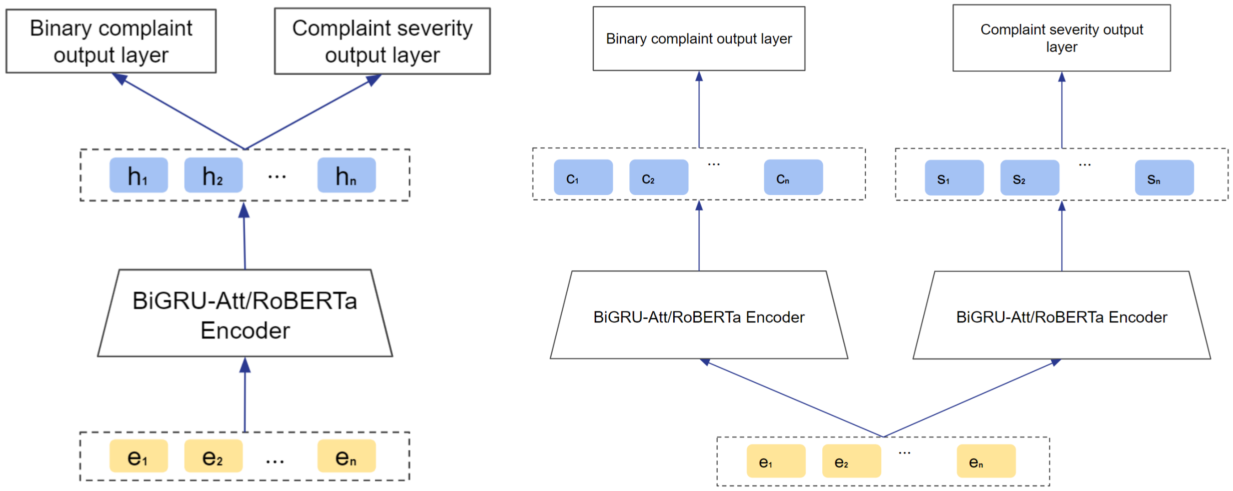
(b) The architecture of MTL-Double Encoder/MTL-Gated Double Encoder.

Figure 4.1: The architecture of multi-task learning models proposed by Rajamanickam et al. (2020).

also concatenate the output of the task-specific and shared encoder and pass it to another BiGRU-Att model for complaint prediction. The rest of the architecture is the same as the MTL-Hard Sharing model (see Figure 4.1b).

MTL-Gated Double Encoder The MTL-Gated Double Encoder model (Rajamanickam et al., 2020) has the same architecture as the MTL-Double Encoder. The outputs from two stacked BiBRU-Att encoders are concatenated by assigning a weight to each representation ($[1-\beta]$ for the output of the task-specific encoder layer and β for the output of the shared one) that controls the importance of the two representations.

MTL-RoBERTa$_{Emo}$ We adapt the best performing model in the severity prediction task (RoBERTa$_{Emo}$) to support multi-task learning by adding an extra output layer for binary complaint prediction (MTL-RoBERTa$_{Emo}$; see Figure 4.2a).


Binary complaint output layer
Complaint severity output layer
h_1
h_2
…
h_n
BiGRU-Att/RoBERTa Encoder
e_1
e_2
…
e_n
Binary complaint output layer
Complaint severity output layer
c_1
c_2
…
c_n
s_1
s_2
…
s_n
BiGRU-Att/RoBERTa Encoder
BiGRU-Att/RoBERTa Encoder
e_1
e_2
…
e_n

(a) The architecture of MTL-BiGRU-Att/MTL-RoBERTa$_{Emo}$.

(b) The architecture of MTL-BiGRU-Att-DE/MTL-RoBERTa$_{Emo}$-DE.

Figure 4.2: The architecture of proposed MTL models based on BiGRU-Att and RoBERTa$_{Emo}$.

MTL-RoBERTa$_{Emo}$-DE We pass the RoBERTa$_{Emo}$ embedding to two separate RoBERTa encoders, i.e. double encoder (DE), followed by two classifiers for binary complaint and severity level prediction (MTL-RoBERTa$_{Emo}$-DE; see Figure 4.2b).

4.3 Experiments Setup

4.3.1 Baselines

We compare the transformer-based models with two previous approaches for complaint identification by Preoţiuc-Pietro et al. (2019) and two neural models.

LR-BOW+Dist. Supervision We present the predictive results from Logistic Regression with bag-of-words trained using the distantly supervised and original complaint data, which has been shown to achieve state-of-the-art results on binary complaint identification (Preoţiuc-Pietro et al., 2019).

LSTM We present the predictive results from a Long-Short Term Memory network (LSTM; Hochreiter and Schmidhuber 1997) that takes as input a tweet, maps its words to embeddings and subsequently passes them through the LSTM to obtain a contextualized representation which is finally fed to the output layer for classification.

ULMFiT We use a transfer learning method, the pre-trained Universal Language Model Fine-tuning model (ULMFiT; Howard and Ruder 2018), for complaint prediction. ULMFiT uses a AWD-LSTM (Merity et al., 2017) encoder for language modeling. It also uses discriminative fine-tuning (tune each layer with different learning rates) and gradual unfreezing (gradually unfreeze the model starting from the last layer) to retain previous knowledge and avoid catastrophic forgetting.

BiGRU-Att We use a standard bidirectional Gated Recurrent Unit (GRU) network (Cho et al., 2014) with a self-attention mechanism (BiGRU-Att; Tian et al. 2018).

MTL-BiGRU-Att & MTL-BiGRU-Att-DE We use two multi-task learning baselines by replacing RoBERTa$_{Emo}$ with BiGRU-Att in the MTL-RoBERTa$_{Emo}$ (MTL-BiGRU-Att; see Figure 4.2a) and MTL-RoBERTa$_{Emo}$-DE models (MTL-BiGRU-Att-DE; see Figure 4.2b).

4.3.2 Hyperparameters

We use **BERT**, **ALBERT** and **RoBERTa** base uncased models and fine-tune them with learning rate l = 1e-5, $l \in$ {1e-4, 1e-5, 2e-5, 1e-6}. We use the base cased pre-trained **XLNet** tuning the learning rate over the same range as for BERT models. For **BERT with linguistic features**, the size of feature embeddings (Emotion, Topics and Emotion+Topics) is $h = 200$, $h \in \{200, 400, 768\}$ with dropout $d = 0.1$, $d \in \{.1, .5\}$ using the same parameters as BERT. We use the default values from Rahman et al. (2020) for the rest of the parameters in BERT with linguistic features. The maximum sequence length is set to 50 covering 95% of tweets in the training set.

For the **MTL-Hard Sharing**, **MTL-Double Encoder** and **MTL-Gated Double Encoder** model, the hidden size of the stacked BiGRU encoder(s) and BiGRU-Att models is h = 128, $h \in$ {64, 128, 256, 512}. We set β in **MTL-Gated Double Encoder** and the remaining parameters in these three models to be the same as Rajamanickam et al. (2020). We train **MTL-RoBERTa**$_{Emo}$ and **MTL-RoBERTa**$_{Emo}$**-DE** with a learning rate $l =$ 1e-6, $l \in \{$1e-5, 5e-6, 1e-6$\}$. The rest of the parameters are the same as RoBERTa$_{Emo}$ in the complaint severity prediction (see Section 3.3.2). The parameter α which controls the importance of the two losses is set to .1, $\alpha \in \{.001, .01, .1, .3, .5\}$.

For **ULMFiT**, we use AWD-LSTM trained on Wikitext-103. We simplify the default fine-tuning by only unfreezing the last 1 layer, the last 2 layers and all layers with learning rates $l_1 = \frac{1e-4}{2.6^4}$, $l_2 = \frac{1e-4}{2.6^3}$ and $l_3 =$ 1e-3 respectively. The **BiGRU-Att**, **MTL-BiGRU-Att** and **MTL-BiGRU-Att-DE** models use 200-dimensional GloVe embeddings (Pennington et al., 2014) pretrained on Twitter data. We train these models using hidden size $h = 128$, $h \in \{64, 128, 256, 512\}$, dropout $d = .2$, $d \in \{.2, .5\}$ and Adam optimizer (Kingma and Ba, 2014) with learning rate $l =$ 1e-3, $l \in \{$1e-3, 5e-3, 1e-2$\}$.

4.3.3 Training and Evaluation

Following Preoţiuc-Pietro et al. (2019), we use a nested 10-fold cross-validation approach as described in Section 3.3.3 to conduct our experiments for complaint prediction. In the outer 10 loops, 9 folds are used for training and one for testing; while in the inner loops, a 3-fold cross-validation method is applied where 2 folds are used for training and one for validation. During training, an early stopping method is applied based on the validation loss. We measure predictive performance using the mean Accuracy, Precision, Recall and macro F1 over 10 folds. We also report the standard deviations.

4.4 Results

Table 4.2 shows the results of STL (top) and MTL (bottom) models on the complaint identification task.[3] Overall, we observe that all **MTL models using M-RoBERTa**$_{Emo}$ perform better than the majority of **STL models**, indicating complaint severity detection

[3] We also conducted preliminary zero-shot experiments using Flan-T5 and ChatGPT (free research preview version) on a small testing set. Results showed that Flan-T5 struggled with accurately identifying binary complaints and detecting severity levels of complaints. ChatGPT demonstrated comparable performance to transformer-based models in detecting binary complaints. However, it fell short in classifying complaints into fine-grained severity categories.

Model	**Acc**	**P**	**R**	**F1**
Single-Task Learning				
Preoţiuc-Pietro et al. 2019				
LR-BOW+Dist. Supervision	81.2	-	-	79.0
LSTM	80.2	-	-	77.0
ULMFiT	82.4 ± 4.5	81.1 ± 4.5	81.8 ± 4.3	81.2 ± 4.5
ULMFiT+Dist. Supervision	83.3 ± 4.7	82.5 ± 4.8	81.8 ± 4.0	81.9 ± 4.6
BiGRU-Att	79.2 ± 5.7	79.2 ± 5.9	74.5 ± 5.5	74.5 ± 5.8
BERT	**88.0** ± 2.9	**87.1** ± 3.3	**87.3** ± 2.8	**87.0** ± 3.0
ALBERT	85.9 ± 2.9	84.8 ± 3.4	84.6 ± 2.9	84.6 ± 3.1
RoBERTa	87.6 ± 3.2	86.6 ± 3.5	86.9 ± 2.9	86.6 ± 3.2
XLNet	83.9 ± 4.1	83.2 ± 4.3	82.3 ± 3.4	82.4 ± 4.0
BERT$_{Emo}$	87.3 ± 3.5	86.5 ± 4.0	86.0 ± 3.7	86.1 ± 3.7
BERT$_{Top}$	87.5 ± 3.3	86.7 ± 3.9	86.5 ± 3.0	86.4 ± 3.4
BERT$_{Emo+Top}$	87.1 ± 2.9	86.4 ± 3.4	85.6 ± 2.7	85.9 ± 2.9
BERT+Dist. Supervision	87.8 ± 3.5	87.0 ± 4.0	86.7 ± 3.3	86.7 ± 3.5
ALBERT+Dist. Supervision	83.9 ± 4.0	82.6 ± 4.3	82.7 ± 3.9	82.6 ± 4.1
RoBERTa+Dist. Supervision	85.2 ± 4.4	84.4 ± 4.7	84.0 ± 3.6	84.0 ± 4.4
XLNet+Dist. Supervision	82.1 ± 4.6	81.7 ± 5.2	79.9 ± 5.2	80.1 ± 4.9
BERT$_{Emo}$+Dist. Supervision	87.7 ± 3.7	86.9 ± 4.3	87.2 ± 3.6	86.8 ± 3.8
BERT$_{Top}$+Dist. Supervision	87.6 ± 4.5	87.0 ± 4.7	86.9 ± 3.8	86.7 ± 4.6
BERT$_{Emo+Top}$+Dist. Supervision	87.8 ± 4.3	87.1 ± 4.7	87.0 ± 3.9	86.9 ± 4.3
Multi-Task Learning				
MTL-BiGRU-Att	77.2 ± 4.9	75.4 ± 4.5	75.7 ± 3.6	75.4† ± 4.5
MTL-BiGRU-Att-DE	75.7 ± 4.8	74.1 ± 4.7	74.6 ± 4.3	74.1 ± 4.7
Rajamanickam et al. 2020				
MTL-Hard Sharing	75.2 ± 4.5	73.5 ± 4.9	71.5 ± 4.4	72.1 ± 4.6
MTL-Double Encoder	74.6 ± 3.5	72.7 ± 3.8	71.7 ± 3.3	72.0 ± 3.6
MTL-Gated Double Encoder	74.7 ± 3.3	73.4 ± 4.1	70.4 ± 2.8	71.1 ± 3.1
MTL-RoBERTa$_{Emo}$	**89.0** ± 3.9	88.2 ± 4.3	**88.4** ± 3.5	**88.2**† ± 4.0
MTL-RoBERTa$_{Emo}$-DE	88.9 ± 3.7	**88.3** ± 4.2	88.3 ± 3.0	88.1 ± 3.7

Table 4.2: Accuracy (Acc), Precision (P), Recall (R) and F1-Score (F1) for complaint identification (± std. dev.). Best results are in bold for STL and MTL respectively. † indicates statistically significant improvement of MTL-BiGRU-Att and MTL-RoBERTa$_{Emo}$ in MTL over BiGRU-Att and BERT in STL respectively (t-test, $p < 0.05$).

improves binary complaint identification.

In *STL settings*, all **transformer-based models** (BERT, ALBERT, RoBERTa and XLNet) perform better than the previous feature-based (LR-BOW+Dist. Supervision) and the non-transformer baselines (ULMFiT, BiGRU-Att), indicating a better capability on capturing idiosyncrasies of complaint syntax and semantics. **BERT** outperforms other models overall across all metrics reaching a macro F1 up to 87, which is 8% higher than the previous state-of-the-art (Preoţiuc-Pietro et al., 2019). The results of **RoBERTa** are close to BERT with 86.6 macro F1 while **ALBERT** and **XLNet** achieve lower performance (84.6 and 82.4 macro F1 respectively). Results of **BERT with linguistic features** are comparable to BERT, among which **BERT**$_{Top}$ is slightly better with 86.4 macro F1. We notice that injecting external linguistic information in BERT's structure for fine-tuning does not help in our case without substantially hurting performance. We speculate that modifying BERT embeddings by injecting extra linguistic information is not complementary to BERT's text representations. Also, **Distant supervision** is beneficial only to **ULMFiT** and **BERT with linguistic features** while BERT and other transformer models perform worse, which is consistent with the results of Bataa and Wu (2019) for sentiment analysis.

In *MTL settings*, **MTL-RoBERTa**$_{Emo}$ outperforms all other models achieving 88.2 macro F1, followed by **MTL-RoBERTa**$_{Emo}$**-DE** with 88.1 F1. This confirms our hypothesis that complaint identification can be benefited from the complaint severity level information when jointly learning these two tasks simultaneously. Also, **MTL-BiGRU-Att** performs better than **BiGRU-Att** in STL achieving 75.4 F1 while the results of **BiGRU-Att** (74.5 F1) and **MTL-BiGRU-Att-DE** (74.1 F1) are comparable. We notice that the **models proposed by Rajamanickam et al. (2020)** (i.e. MTL-Hard sharing, MTL-Double Encoder and MTL-Gated Double Encoder) achieve low performance with only the MTL-Hard Sharing model performing slightly better than the others with 72.1 macro F1. We speculate that adding one or more extra BiGRU encoders before the BiGRU-Att model is an overly complex structure for our data set.

4.5 Discussion

4.5.1 Analysis in Single-Task Learning

We investigate the limitations in predicting capacity of the best performing model in single-task learning settings (i.e. **BERT**). We randomly analyze 100 cases in predictive results, where 50 cases were misclassified as non-complaints and another 50 cases were misclassified as complaints.

In cases where **complaints** were misclassified as **non-complaints**, 26% errors are due to implicit expressions while 14% errors are because complaints contain irony. In the former situation, complaints express weak emotional intensity without explicit reproach, where complainers imply their dissatisfaction instead of directly complaining or mentioning the cause (Trosborg, 2011). The following tweet is a typical example:

> T1: *"It started yesterday, but I try again it could work normal. But since last night its just like this <url>"*

Such expressions rarely include words related to complaints (e.g. *"disappointed"*, *"bad service"*) and are therefore difficult to be correctly classified. In the latter situation, complaints are expressed in an ironic way using terms such as *"congratulations"*, *"thank you"* and *"brilliant"*. For instance, the following text was wrongly classified as a non-complaint:

> T2: *"Thank you so much for making a box that shreds apart even when carried by both handles."*

In cases where **non-complaints** were misclassified as **complaints**, errors can be roughly divided into four categories: (1) 26% errors are because certain terms appear frequently in complaints during training such as *"thank you"*, *"dm"*, *"lost"*, *"work"*. The following non-complaint was wrongly classified as a complaint:

> T3: *"BTW <user> – <user> did me right, and replaced my two failed batteries under warranty. I'm* **happy** *:)* **thanks** *<user>!"*

It contains similar words with the following complaint in the same fold (similarities highlighted in bold):

T4: *"Was* **happy** *to find out <user> had an app to watch all their shows, until 6 episodes in it stops working.* **Thanks**! *<user>"*

(2) 22% errors due to interrogative tone, which is common in complaints. An example is *"Folks, what is cost of text message to a us number?"* (3) 22% errors are from negation words such as *"No luck with pc or phone."* (4) 12% errors are because texts contain negative sentiment such as *"This would be a terrible idea <url>"* are likely to be classified as complaints incorrectly since words such as *"terrible"* are widely used to express dissatisfaction. However, there are not enough cues to indicate a violation of expectations. According to the statistics, the proportion of complaints misclassified as non-complaints (15.22%) is higher than that of non-complaints misclassified as complaints (10.25%) indicating implicit and figurative expressions as well as unknown factors in complaints are more challenging to identify.

4.5.2 Analysis in Multi-Task Learning

We also investigate the influence of recognizing severity levels of complaints on binary complaint identification in our MTL settings. We analyze predictive results by inspecting predictions from the best performing model in STL (i.e. **BERT**) and **MTL-RoBERTa**$_{Emo}$ model in a random fold (out of 10 cross validation folds). We observe that 9.8% of predictions flip, where the number of complaints flipping to non-complaints is noticeably larger (88.2%) than that of non-complaints flipping to complaints (11.8%). Similarly, we also compare predicted results between **BiGRU-Att** (STL) and **MTL-BiGRU-Att** in the same fold. The flipping percentage (6.9%) is lower than BERT and MTL-RoBERTa$_{Emo}$ while the proportions of one class flipping to another are consistent (83.4% and 16.6% respectively). These indicate that complaint severity information encapsulates complementary information for the model to predict non-complaints accurately.

Table 4.3 shows flipping examples from **BERT** (STL) and **MTL-RoBERTa**$_{Emo}$. From the first two rows, we see that the MTL model is not affected by negation (e.g. *"never"*) and negative terms (e.g. *"bad"*, *"very low"*) using the extra knowledge provided by the severity level prediction task. Also, in the last two examples, complaints are expressed in a more subtle way that rarely contains typical complaint-related terms. This indicates that the MTL model is able to detect this type of complaints correctly because the severity level information encourages the model to learn to distinguish between such stylistic idiosyncrasies.

Tweet	**BERT**	**MTL-RoBERTa$_{Emo}$**	Actual Label	Severity Label
What's your secret to poaching eggs? Mine **never** *look that good.*	Complaint	Non-complaint	Non-complaint	No Complaint Severity
<URL> How **bad** *do you really want a ps4 this year? Get a pre-owned playstation 4 at a* **very low** *dis <URL>*	Complaint	Non-complaint	Non-complaint	No Complaint Severity
So, I'm now having to check my <USER> forester's oil each month. Put 4 quarts in today, got about 2 out. #smh	Non-complaint	Complaint	Complaint	Disapproval
ls this how you fix the exhaust of your <USER> in #belarus? <URL>	Non-complaint	Complaint	Complaint	Blame

Table 4.3: Complaint classification examples by BERT and MTL-RoBERTa$_{Emo}$ compared to the actual labels.

We further observe that 11.2% of wrong predictions remain the same, where complaints and non-complaints account for 59.0% and 41.0% respectively which means severity features benefit more posts that are complaints to be classified accurately. However, the model still has difficulty in predicting some posts which might happen because of the lower performance of severity detection[4] when used as an auxiliary task in the MTL settings.

4.6 Cross Domain Experiments

Finally, we use BERT to train models on one domain and test on another as well as training on all domains except the one that the model is tested on. Table 4.4 shows the performance of the model in Preoţiuc-Pietro et al. (2019) (LR-BOW+Dist. Supervision on the left) and BERT (right) across 9 domains.

We first observe that BERT results in the scores of nearly half of the cases are lower than the results of LR-BOW+Dist. Supervision when training on a single domain (especially “Food’, “Car” and “Other”) while BERT trained on all domains performs better across all testing domains, achieving a macro F1 up to 88.2 when tested on “Other”. This indicates that fine-tuning BERT on a small training data set (“Food”, “Car” and “Other” are three of the domains with the smaller amount of data) is not enough to make it perform well.

[4] Severity prediction is less accurate in a MTL than in a STL.

Test **Train**	F		A		R		C		Se		So		T		E		O	
Food	-		**69.7**	49.8	**76.5**	53.2	**85.7**	61.8	**73.3**	56.8	**76.2**	59.2	**71.2**	52.7	**74.4**	61.1	**83.0**	48.7
Apparel	**75.0**	69.7	-		74.2	**81.5**	**78.8**	74.3	72.7	**76.3**	73.2	**81.8**	66.7	**74.1**	75.7	**75.9**	81.6	**84.2**
Retail	74.9	**75.9**	72.8	**80.0**	-		**80.0**	73.0	75.4	**75.9**	75.9	**80.7**	70.0	**74.9**	72.7	**75.3**	**86.6**	79.5
Cars	**76.1**	57.1	**70.2**	62.1	**75.5**	65.1	-		**70.2**	51.6	**75.1**	70.6	**71.3**	62.3	**74.4**	61.4	**82.1**	71.8
Services	**79.8**	64.7	71.1	**82.4**	**77.2**	76.5	**77.5**	75.4	-		73.8	**78.6**	73.4	**76.3**	75.5	**79.4**	78.9	**83.0**
Software	**74.6**	69.3	70.4	**80.0**	73.5	**77.5**	**79.1**	78.0	**77.9**	76.4	-		73.4	**75.2**	**76.7**	76.2	81.7	**82.0**
Transport	**72.5**	62.2	70.5	**73.5**	77.1	**80.0**	**80.0**	79.2	74.4	**76.3**	**75.8**	75.3	-		**72.4**	70.4	82.0	**82.6**
Electronics	69.9	**72.9**	72.2	**78.5**	69.1	**78.9**	73.0	**75.4**	77.0	**78.0**	71.0	**72.4**	**69.8**	69.7	-		**82.1**	80.8
Other	**65.9**	64.8	**75.2**	74.8	**79.2**	72.2	**81.7**	69.2	**76.3**	69.9	76.5	**77.9**	70.6	70.6	**72.8**	70.8	-	
All	48.9	**77.5**	67.5	**87.7**	72.6	**85.8**	73.9	**80.9**	72.0	**81.1**	65.8	**85.1**	64.9	**81.4**	67.6	**82.0**	81.9	**88.2**

Table 4.4: F1-score of models in Preoţiuc-Pietro et al. (2019) (left) and BERT (right) trained from one domain and tested on other domains. Domains include Food (F), Apparel (A), Retail (R), Cars (C), Services (Se), Software (So), Transport (T), Electronics (E) and Other (O). The **All** line shows results on training on all categories except the category in testing. Best results are in bold.

In contrast, it achieves better performance consistently on larger data sets (All). We also notice that BERT performs robustly for domain pairs where the domains are either used for training or testing. For example, training on "Apparel" achieves high performance when testing on "Software" (81.8 F1) and vice versa (80.0 F1). Furthermore, domain relevance affects predictive performance. For example, BERT trained on "Transport" achieves 79.2 F1 when tested on "Car", which is the highest performance compared to other training domains since these two domains share common vocabulary (see "Car" column for BERT).

4.7 Summary

In this chapter, we evaluated different transformer-based models and their combinations with linguistic information on complaint identification. We proposed MTL settings which jointly model complaint identification (main task) and complaint severity level classification (auxiliary task) to boost the predictive performance. The results showed that the complaint severity is beneficial to binary complaint prediction. We also performed an error analysis and found that models struggled with implicit and ironic expressions. Furthermore, we conducted cross domain experiments to explore the domain adaptability of predictive models.

Chapter 5

Automatic Identification and Classification of Bragging

In Section 2.3, bragging is a speech act employed with the goal of constructing a favorable self-image through positive statements about oneself and their close networks (Dayter, 2014, 2018). It is widespread in daily communication and especially popular in social media, where users aim to build a positive image of their persona directly or indirectly (Ren and Guo, 2020)

However, bragging is considered a high risk act (Brown and Levinson, 1987; Holtgraves, 1990; Van Damme et al., 2017) and can lead to the opposite effect than intended, such as dislike or decreased perceived competence (Jones et al., 1982; Sezer et al., 2018; Matley, 2018) (see Section 2.3). It is, thus, paramount to understand the types of bragging and strategies to mitigate the face threat introduced by bragging as well as how effective the self-presentation attempt is (Herbert, 1990).

Although bragging has aroused great interest in linguistics and psychology (see Section 2.3.2), it has yet to be studied at scale in computational (socio) linguistics. The ability to identify bragging automatically is important for: (a) linguists to better understand the context and types of bragging through empirical studies (Dayter, 2014; Ren and Guo, 2020); (b) social scientists to analyze the relationship between bragging and personality traits, online behavior and communication strategies (Miller et al., 1992; Van Damme et al., 2017; Sezer et al., 2018); (c) online users to enhance their self-presentation strategies (Miller et al., 1992; Dayter, 2018); (d) enhancing NLP applications such as intent identification (Wen et al.,

Type	**Definition**	**Tweet**
Achievement	Concrete outcome obtained as a result of the tweet author's actions. These may include accomplished goals, awards and/or positive change in a situation or status (individually or as part of a group).	*Finally got the offer! Whoop!!*
Action	Past, current or upcoming action of the user that does not have a concrete outcome.	*Guess what! I met Matt Damon today!*
Feeling	Feeling that is expressed by the user for a particular situation.	*Im so excited that I am back on my consistent schedule. I am so excited for a routine so I can achieve my goals!!*
Trait	A personal trait, skill or ability of the user.	*To be honest, I have a better memory than my siblings*
Possession	A tangible object belonging to the user.	*Look at our Christmas tree! I kinda just wanna keep it up all year!*
Affiliation	Being part of a group (e.g. family, fanclub, university, team, company, etc.) and/or a certain location including living in a city, neighborhood or country.	*My daughter got first place in the final exam, so proud of her!*
Not Bragging	The tweet is not about bragging or (a) there is not enough information to determine that the tweet is about bragging; (b) the bragging statements belong to someone other than the author of the tweet; (c) the relationship between the author and people or things mentioned in the tweet are unknown.	*Glad to hear that! Well done Jim!*

Table 5.1: Bragging taxonomy together with type definitions and examples of tweets.

2017) and conversation modeling (Lin et al., 2020).

The main contributions of this chapter are as follows: (1) we create a new publicly available data set containing a total of 6,696 English tweets annotated with bragging and their types; (2) we experiment with transformer-based models (Vaswani et al., 2017) that inject linguistic information for bragging identification (binary classification) and bragging type classification (seven classes); (3) we present a qualitative linguistic analysis of markers of bragging in tweets and the model behavior in predicting bragging.

The work presented in this chapter has been published at the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) (Jin et al., 2022).

5.1 Data

5.1.1 Bragging Definition & Types

Definition In Section 2.3, bragging has been defined as a speech act that explicitly or implicitly attributes credit to the speaker for some *good* (e.g. possession, skill) that is positively valued by the speaker and their audience (Dayter, 2014). A bragging statement should clearly express what the speaker is bragging about.

Types We generalize and extend the bragging types based on the definitions by Dayter (2018) and Matley (2018). The former summarizes them as accomplishments and some aspects of self; while the latter includes everyday achievements (e.g. cooking) and personal qualities. We divide the "some aspects of self" category into two categories, namely "Possession" and "Trait" respectively. We also add an "Affiliation" category for bragging involving a group to which the speaker belongs. In total, we consider six bragging types and a non-bragging category. Table 5.1 shows the definitions and examples of each bragging type.

Classification Tasks Given the taxonomy above, we define two classification tasks: (i) *binary* bragging prediction (i.e. if a tweet contains a bragging statement or not); and (ii) seven-way *multiclass* classification for predicting if a tweet contains one of the six bragging types or no bragging at all.

5.1.2 Data Collection

To the best of our knowledge, there is no other data set available for our study. We use Twitter for data collection as tweets are openly available for research and widely used in other related tasks, e.g. predicting sentiment (Rosenthal et al., 2017), affect (Mohammad et al., 2018), sarcasm (Bamman and Smith, 2015), stance (Mohammad et al., 2016).

Random Sampling We select tweets for annotation by randomly sampling from the 1% Twitter feed one day per month from January 2019 to December 2020 (approximately 10k tweets per day) to ensure diversity using the Premium Twitter Search API for academic research.[1]

Keyword-based Sampling To give a model access to more positive examples of bragging statements for training, we use a keyword-based sampling method that increases the hit rate of bragging, following previous work on labeling infrequent linguistic phenomena, e.g. irony (Mohammad et al., 2018) or hate speech (Waseem and Hovy, 2016).

We build queries based on indicators of positive self-disclosure (e.g. *I*, *just*) (Dayter, 2018) and stylistic indicators, e.g. positive emotion words, present tense verbs (Bazarova et al., 2013). As the frequency of these keywords is high, we construct multi-word queries consisting of a personal pronoun and an indicator. In addition, we use a short list of curated bragging-related hashtags. The queries are: {[*I*, *proud*], [*I*, *glad*], [*I*, *happy*], [*I*, *best*], [*I*, *amazed*], [*I*, *amazing*], [*I*, *excellent*], [*I*, *just*], [*I'm*, *proud]*, [*I'm*, *glad*], [*I'm*, *happy*], [*I'm*, *best*], [*I'm*, *amazed*], [*I'm*, *amazing*], [*I'm*, *excellent*], [*me*, *proud*], [*my*, *best*], *#brag*, *#bragging*, *#humblebrag*, *#humble*, *#braggingrights*}. After annotating 1,000 tweets, we compute the percentage of bragging tweets for each keyword and remove from sampling tweets with less than 5% (i.e. [*I*, *amazed*], [*I'm*, *amazing*], [*I'm*, *best*], [*my*, *best*], [*I*, *excellent*], *#humble*).

We initially collected around 6K and 368K tweets using hashtags and multi-word queries respectively. We obtain over 9k tweets by keeping all tweets collected using hashtags and sampling 1% from those collected using multi-word queries to balance the two types.

Data Filtering After collecting tweets, we exclude those with duplicate or no meaningful textual content (e.g. only @-mentions or images). We only focus on English posts and filter

[1] https://tinyurl.com/2p8wnure

Data set	Precision	Recall	Macro-F1
Single Annotation	70.99±1.33	68.37±1.13	69.53±0.97
Multiple Annotations	70.66±0.64	68.49±0.83	69.47±0.32

Table 5.2: Precision, Recall and macro F1-Score obtained for binary bragging classification using the same testing set annotated by a single annotator and by multiple annotators.

out non-English ones using the language code provided by Twitter. We also exclude retweets and quoted tweets, as these do not typically express the thoughts of the user who retweet them. Moreover, we exclude 131 tweets containing a URL in the text because these are related to advertisements based on initial results from our annotation calibration rounds. This resulted in a total of 6,696 tweets which is of similar size to data sets recently released for social NLP (Oprea and Magdy, 2020; Chung et al., 2019; Beck et al., 2021; Mendelsohn et al., 2021).

5.1.3 Annotation and Quality Control Process

We manually annotate tweets for providing a solid benchmark and foster future research. Annotators are four authors of this work (Jin et al., 2022) consisting of 2 females and 2 males/1 PhD student, 2 academic staff and 1 industry staff. All annotators have significant experience in linguistic annotation. We run three calibration rounds of 100 tweets each, where all annotate all tweets and discuss disagreements until a Krippendorf's Alpha above 0.80 in the seven-class task is reached.

To monitor quality, a subset of 1,564 tweets were annotated by two annotators or more in case of disagreements. If a tweet fits into multiple bragging types, we assign the more prominent one. The annotation is based only on the actual text of the tweet without considering additional modalities (e.g. images), context or replies. This is similar to the information available to predictive models during training. We select the final label as the majority vote and a final label was assigned after consensus in cases of two different votes. The full task guideline, examples and interface are presented in Appendix B.

Quality of a Single Annotator We conduct two tests to assess the quality of the data set annotated by a single annotator. Firstly, we evaluate the binary performance of BERT on a subset annotated by multiple annotators and by a single annotator chosen randomly from

Label	**Training set**	**Dev/Test set**	**All**
	(Keyword sampling)	**(Random sampling)**	
Binary			
Bragging	544 (16.09%)	237 (7.15%)	781 (11.66%)
Not Bragging	2,838 (83.91%)	3,077 (92.85%)	5,915 (88.34%)
Multi-class			
Achievement	166 (4.91%)	71 (2.14%)	237 (3.54%)
Action	127 (3.76%)	58 (1.72%)	185 (2.76%)
Feeling	39 (1.15%)	27 (0.82%)	66 (0.99%)
Trait	91 (2.69%)	48 (1.45%)	139 (2.08%)
Possession	58 (1.72%)	28 (0.84%)	86 (1.28%)
Affiliation	63 (1.86%)	5 (0.15%)	68 (1.01%)
Not Bragging	2,838 (83.91%)	3,077 (92.85%)	5,915 (88.34%)
Total	3,382	3,314	6,696

Table 5.3: Bragging data set statistics.

them. Labels in the two training sets are from these two types of annotators separately while labels in the testing sets are the same (i.e. both are annotated by multiple annotators). The results in Table 5.2 show single annotations do not lead to a significant drop across all metrics. Secondly, we manually evaluate a batch of 100 labels annotated by a single annotator via annotating by another annotator. The two annotators agree on 93 annotations out of 100. Then a third annotator annotates these divisive labels and agrees with 2 annotations by the original annotator and disagrees with 1 annotation by both annotators. This means that a single annotator and multiple annotators achieve consensus on 95 out of 99 annotations.

The inter-annotator agreement between two annotations of all tweets is (a) percentage agreement: 89.03; (b) Krippendorff's Alpha (Krippendorff, 2011) (7-class): 0.840; (c) Krippendorff's Alpha (binary): 0.786. Agreement values are between the upper part of the *substantial* agreement band and the *perfect* agreement band (Artstein and Poesio, 2008). The final data set consists of 6,696 tweets with one of the seven classes. Before annotation, the keyword-based and randomly sampled tweets were shuffled to not induce frequency bias. Data set statistics are shown in Table 5.3, including statistics across the two sampling strategies. The model performance curve by varying the training set size indicates that annotating more data is not likely to lead to substantial improvements in bragging prediction (see Figure 5.1).

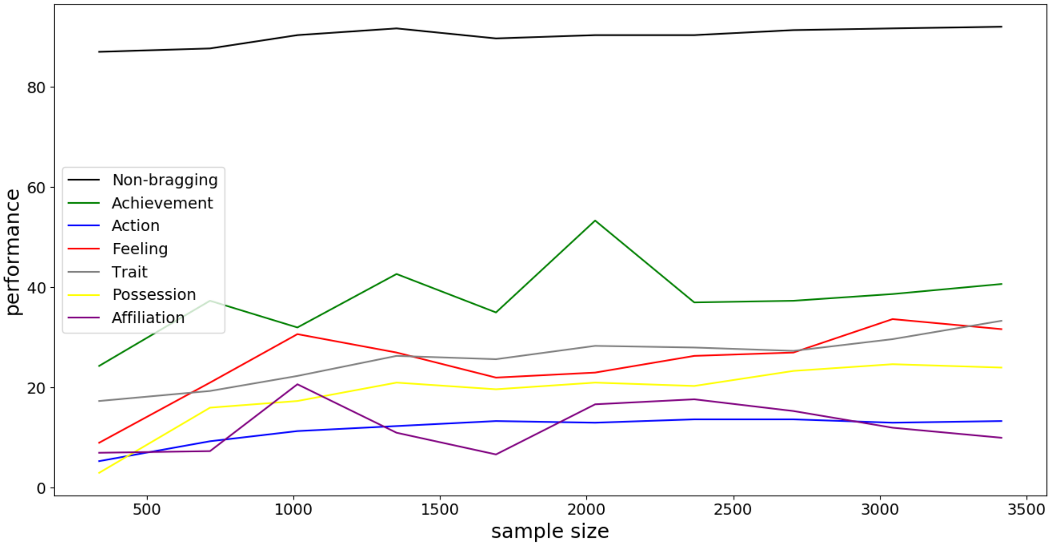


Figure 5.1: Learning curve for performance across each bragging type.

Class	Self-disclosure (%)	Non-self-disclosure (%)
Bragging	31.63	68.37
Non-bragging	24.04	75.96
Achievement	31.65	68.35
Action	27.57	72.43
Feeling	31.82	68.18
Trait	36.69	63.31
Possession	29.07	70.93
Affiliation	35.29	64.71
Non-bragging	24.04	75.96
Total	24.93	75.07

Table 5.4: Percentages of self-disclosure class across bragging classes.

5.1.4 Self-Disclosure in Bragging

We conduct an analysis of the relationship between self-disclosure and bragging as they are closely related. We use the self-disclosure lexicon by Bak et al. (2014) to assign each tweet in our data set a label (i.e. self-disclosure or non-self-disclosure). The percentages of self-disclosure across each bragging type are shown in Table 5.4. We also use self-disclosure models as a predictor for bragging in early experimentation but the results are omitted due to the low performance.

5.1.5 Data Splits

We use the keyword sampled data for training and the random data for development and testing (in the ratio of 2:8) because the latter is representative of the real distribution of tweets (see Table 5.3).

5.2 Methodology

We evaluate vanilla transformer-based models (Vaswani et al., 2017) and further leverage external linguistic information to improve them.

5.2.1 BERT, RoBERTa and BERTweet

We experiment with Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2019), RoBERTa (Liu et al., 2019) and BERTweet (Nguyen et al., 2020). RoBERTa is a more robust variant of BERT that obtains better results on a wide range of tasks (Liu et al., 2019). BERTweet is pre-trained on English tweets using RoBERTa as the basis and achieves better performance on Twitter tasks (Nguyen et al., 2020). We fine-tune BERT, RoBERTa and BERTweet for binary and multiclass bragging prediction by adding a classification layer that takes the [CLS] token as input.

5.2.2 BERTweet with Linguistic Features

We inject linguistic knowledge that could be related to bragging to the BERTweet model with a similar method described in Section 3.2.2,[2] that was found to be effective on complaint severity classification (see Section 3.4), a related pragmatics task. The method is adapted from Rahman et al. (2020), which integrates multimodal information (e.g. audio, visual) in transformers using a fusion mechanism called Multimodal Adaption Gate (MAG). MAG integrates multimodal information to text representations in transformer layers using an attention gating mechanism for modality influence controlling. We first expand vectors of linguistic features to a comparable size to the BERTweet embeddings. Then, we use MAG

[2] Early experimentation with simply concatenating or applying attention resulted in lower performance.

to concatenate contextual and linguistic representations after the embedding layer of the transformer similar to Rahman et al. (2020). The output is sent to a pre-trained BERTweet encoder for fine-tuning followed by an output layer.

We experiment with these linguistic features:

- **NRC:** The NRC word-emotion lexicon contains a list of English words mapped to ten categories related to emotions and sentiment (Mohammad and Turney, 2013). We represent each tweet as a 10-dimensional vector where each element is the proportion of tokens belonging to each category.
- **LIWC:** Linguistic Inquiry and Word Count (Pennebaker et al., 2001) is a dictionary-based approach to count words in linguistic, psychological and topical categories. We use LIWC 2015 to represent each tweet as a 93-dimensional vector.
- **Clusters:** We use Word2Vec clusters proposed by Preoţiuc-Pietro et al. (2015b) to represent each tweet as a 200-dimensional vector over thematic subjects.

5.3 Experimental Setup

5.3.1 Text Processing

We pre-process text by lowercasing, replacing all username mentions with placeholder tokens *@USER* and emojis with words using demojize.[3] We also remove hashtags that are used as keywords (e.g. *#brag*) in data collection. Finally, we tokenize the text using TweetTokenizer.[4]

5.3.2 Baselines

Majority Class: As a first baseline, we label all tweets with the label of the majority class.

[3] https://pypi.org/project/emoji/

[4] https://www.nltk.org/api/nltk.tokenize.html

LR-BOW: We train a Logistic Regression with bag-of-words using L2 regularization.

BiGRU-Att: We also train a bidirectional Gated Recurrent Unit (GRU) network (Cho et al., 2014) with self-attention (BiGRU-Att; Tian et al. 2018). Tokens are first mapped to GloVe embeddings (Pennington et al., 2014) and then passed to a bidirectional GRU. Subsequently, its output is passed to a self-attention layer and an output layer for classification.

5.3.3 Hyperparameters

For **BiGRU-Att**, we use 200-dimensional GloVe embeddings (Pennington et al., 2014) pre-trained on Twitter data. The hidden size is $h = 128$ where $h \in \{64, 128, 256, 512\}$ with dropout $d = .2$, $d \in \{.2, .5\}$. We use Adam optimizer (Kingma and Ba, 2014) with learning rate $l =$ 1e-2, $l \in \{$1e-3, 1e-2, 1e-1$\}$. For **BERT**, **RoBERTa** and **BERTweet**, we use the base cased model (12 layers and 109M parameters, 12 layers and 125M parameters and 12 layers and 135M parameters accordingly) and fine-tune them with learning rate $l =$ 3e-6, $l \in \{$1e-4, 1e-5, 5e-6, 3e-6, 1e-6$\}$. For **BERTweet with linguistic features**, we project these to vectors of size $l_{NRC} = 200$, $l_{LIWC} = 400$, $l_{Clusters} = 768$, $l \in \{10, 93, 200, 400, 600, 768\}$. For MAG, we use the default parameters from Rahman et al. (2020). For **multi-class classification**, we apply class weighting due to the imbalanced data and set the training epoch to $n = 40$, $n \in \{15, 20, 25, 30, 35, 40, 45, 50, 55, 60\}$. The maximum sequence length is set to 50 covering 95% of tweets in the training set. We use a batch size of 32.

5.3.4 Training and Evaluation

We train each model three times using different random seeds and report the mean Precision, Recall and F1 (macro). We apply early stopping during training based on the dev loss. The experiments with linguistic features are performed with the best pre-trained transformer in each of the two classification tasks.

Model	Precision	Recall	Macro-F1	Precision	Recall	Macro-F1
	Bragging Classification (Binary)			Bragging and Type Classification (7 class)		
Majority Class	46.42	50.00	48.15	13.26	14.29	13.76
LR-BOW	54.53	63.16	52.68	18.52	20.02	18.59
BiGRU-Att	55.93 ± 1.53	51.41 ± 0.47	51.29 ± 1.40	18.32 ± 0.10	26.16 ± 3.41	19.19 ± 0.31
BERT	64.24 ± 1.40	65.91 ± 3.32	64.58 ± 0.80	24.16 ± 1.15	39.66 ± 4.84	26.85 ± 0.81
RoBERTa	66.53 ± 0.29	68.43 ± 2.05	67.34 ± 1.02	28.99 ± 0.61	45.90 ± 3.59	32.82 ± 0.65
BERTweet	70.43 ± 0.16	**72.62** ± 0.89	71.44 ± 0.43	30.82 ± 0.75	47.25 ± 2.68	34.86 ± 0.79
BERTweet-NRC	72.89 ± 1.26	70.95 ± 0.96	71.80 ± 0.49	30.95 ± 0.54	**47.98** ± 1.12	34.36 ± 0.19
BERTweet-LIWC	72.65 ± 0.20	72.21 ± 0.43	**72.42**† ± 0.31	32.06 ± 2.42	46.68 ± 7.45	34.83 ± 0.79
BERTweet-Clusters	71.26 ± 2.27	72.53 ± 1.91	71.60 ± 0.21	**32.51** ± 1.36	46.97 ± 2.36	**35.95** ± 0.54

Table 5.5: Macro precision, recall and F1-Score (± std. dev. for 3 runs) for bragging prediction (binary and multiclass). Best results are in bold. † indicates significant improvement over BERTweet (t-test, $p<0.05$).

5.4 Results

5.4.1 Binary Bragging Classification

Table 5.5 (left) shows the predictive performance of all models on predicting bragging (i.e. binary classification). Overall, BERTweet models with linguistic information achieve better overall performance. **Transformer** models perform substantially above the **majority class** baseline (+23.29 F1) and above **Logistic Regression** (+18.76 F1). **BERTweet** (71.44 F1) performs better than **BERT** (64.58 F1) and **RoBERTa** (67.34 F1), which illustrates the advantage of pre-training on English tweets for this task.

Performance is further improved (+0.98 F1) by using *LIWC features* alongside **BERTweet**, which indicates that injecting extra linguistic information benefits bragging identification. We speculate that this is because a bragging statement usually contains particular terms (e.g. personal pronouns, positive terms) or involves at least one certain aspect or theme (e.g. reward or property), which can be captured by linguistic features (e.g. feature *I* and *ACHIEVE* in LIWC). Combining lexicons leads to worse results than using a single one, so we refrain from reporting these results for clarity.

5.4.2 Multi-class Bragging Classification

Table 5.5 (right) shows the predictive performance of all models on multiclass bragging type prediction including not bragging. We again find that pre-trained **transformers** substantially outperform the **majority class** baseline (+21.1 F1) and **logistic regression** (+16.27 F1). In line with the binary results, we find that **BERTweet** (34.86 F1) performs best out of all transformers. **BERTweet-Clusters** outperforms all models (35.95 F1), which indicates that topical information helps to identify different types of bragging. Each bragging type might be particularly specialized to certain topics (e.g. weight loss in "Achievement" category).

5.5 Discussion

5.5.1 Linguistic Feature Analysis

We analyze the linguistic features, i.e. unigrams, LIWC and part-of speech (POS) tags associated with bragging and its type in all tweets of our data set. For this purpose, we first tag all tweets using the Twitter POS Tagger (Derczynski et al., 2013). Each tweet is represented as a bag-of-words distribution over POS unigrams and bigrams to reveal distinctive syntactic patterns of bragging and its type. For each unigram, LIWC and POS feature, we compute correlations between its distribution across posts and the label of the post. Then, we use the method introduced by Schwartz et al. (2013) to rank the features using univariate Pearson correlation with words normalized to sum up to unit for each tweet.

Table 5.6 presents the top 15 features from unigrams (lowercase) and LIWC (uppercase) and the top 10 features from POS unigrams and bigrams correlated with bragging and non-bragging tweets. We notice that the top words in the bragging category can be classified into (a) personal pronouns (e.g. *my*, *I*) that usually indicate the author of the bragging statement; (b) words related to time (e.g. *FOCUSPAST*, *TIME*, *during*); and (c) words related to a specific bragging target (e.g. *RELATIV*, *ACHIEVE*, *REWARD*, *managed*). These findings are in line with the indicators of positive self-disclosure by Dayter (2018) and Bazarova et al. (2013). Furthermore, personal pronouns followed by a verb in the past tense (i.e. *PRP_VBD*) is common in bragging (e.g. *"I forgot what it's like to be good at school. Today I finished a thing we were doing so fast that everyone around me started asking ME*

Bragging		**Non-Bragging**	
Feature	r	Feature	r
Unigrams (lowercase) and LIWC (uppercase)			
AUTHENTIC	0.149	CLOUT	0.109
my	0.127	YOU	0.089
I	0.122	DISCREP	0.078
TONE	0.104	NEGEMO	0.077
FOCUSPAST	0.102	SOCIAL	0.076
WC	0.100	FOCUSPRESENT	0.070
RELATIV	0.090	INFORMAL	0.056
TIME	0.081	COGPROC	0.056
during	0.078	ANGER	0.056
ACHIEVE	0.075	just	0.054
PREP	0.073	your	0.052
managed	0.072	IPRON	0.051
REWARD	0.069	?	0.043
row	0.068	not	0.038
got	0.067	why	0.037
POS (Unigrams and Bigrams)			
PRP_VBD	0.104	NNP	0.081
VBD	0.093	VB	0.061
CD_NNS	0.077	RB_VB	0.056
PRP$	0.074	NNP_NNP	0.049
VBD_DT	0.062	VBP_PRP	0.048
NN_IN	0.061	VBZ	0.039
IN_CD	0.060	MD	0.035
IN_PRP$	0.060	NNP_VBZ	0.033
PRP$_NN	0.058	RB_RB	0.031
VBD_PRP$	0.057	MD_PRP	0.031

Table 5.6: Feature correlations including unigrams (lowercase), LIWC (uppercase), part-of-speech (POS) unigrams and bigrams with bragging and non-bragging tweets, sorted by Pearson correlation (r). All correlations are significant at $p < .01$, two-tailed t-test.

Achievement		**Action**		**Feeling**		**Trait**		**Possession**		**Affiliation**	
Feature	r	Feature	r	Feature	r	Feature	r	Feature	r	Feature	r
Unigrams (lowercase) and LIWC (uppercase)											
FOCUSPAST	0.200	get	0.146	happy	0.228	APOSTRO	0.197	own	0.211	FAMILY	0.276
Number	0.157	trip	0.128	POSEMOE	0.218	COGPROC	0.181	buy	0.175	CLOUT	0.271
Analytic	0.153	RELATIV	0.119	❤️	0.191	FOCUSPRESENT	0.179	bought	0.149	proud	0.263
finished	0.150	ready	0.114	blessed	0.190	cute	0.159	car	0.146	rights	0.215
3	0.133	him	0.114	AFFECT	0.184	PRONOUN	0.157	bedroom	0.144	SOCIAL	0.209
WORK	0.132	happen	0.105	feels	0.176	take	0.143	extra	0.144	amazing	0.205
managed	0.130	FOCUSFUTURE	0.105	love	0.169	COMPARE	0.143	xr	0.142	💙	0.197
over	0.129	fun	0.102	sunrise	0.166	ANGER	0.138	macbook	0.055	law	0.185
under	0.119	gave	0.097	weighted	0.162	I	0.137	new	0.139	team	0.182
beat	0.112	hours	0.096	july	0.159	if	0.137	afford	0.139	OTHERP	0.181
race	0.104	before	0.095	time	0.159	SWEAR	0.134	PERIOD	0.106	words	0.164
office	0.103	sitting	0.095	truly	0.156	am	0.133	HOME	0.105	teams	0.164
possible	0.103	VERB	0.094	BIO	0.147	PPRON	0.132	DASH	0.084	#baseball	0.164
5	0.101	PREP	0.089	CERTAIN	0.143	me	0.130	I	0.077	fan	0.163
SIXLTR	0.100	INGEST	0.085	TONE	0.140	look	0.122	DISCREP	0.071	MALE	0.160
POS (Unigrams and Bigrams)											
CD_NNS	0.198	DT_NNP	0.139	RB_JJ	0.183	VBP	0.252	$_CD	0.161	FW_,	0.164
VBD	0.171	VBP_TO	0.124	VBP_IN	0.174	PRP	0.193	$	0.130	VB_VBD	0.161
CD	0.164	IN_:	0.117	VB_RBR	0.161	PRP_VBP	0.191	NN_PDT	0.130	CC_UH	0.159
NNS	0.145	VBP_WP	0.116	JJR_WRB	0.161	VBP_JJ	0.162	NNS_UH	0.122	VBZ_DT	0.151
VBD_DT	0.141	NNP_UH	0.116	RB_VBZ	0.146	UH_DT	0.150	SYM_:	0.114	DT_RBS	0.146
PRP_VBD	0.132	NFP_NNP	0.116	CC_JJ	0.143	VBP_DT	0.150	VBZ_JJ	0.110	UH_NNP	0.145
NN_IN	0.132	NNP	0.116	VBD_:	0.131	RB_VB	0.149	VB_PRP$	0.109	._SYM	0.138
IN_CD	0.130	NNP_NNS	0.114	._VBG	0.123	MD	0.149	PRP$_JJ	0.109	NFP_CC	0.137
VBN	0.129	TO_VB	0.109	UH_WP	0.118	MD_VB	0.134	._VBD	0.109	PRP_PRP$	0.136
VB_JJR	0.109	TO	0.107	POS_RB	0.118	CC_WP	0.131	NN_PRP$	0.106	NN_NN	0.135

Table 5.7: Feature correlations including unigrams (lowercase), LIWC (uppercase), part-of-speech (POS) unigrams and bigrams with tweets containing six bragging types, sorted by Pearson correlation (r). All correlations are significant at $p < .01$, two-tailed t-test.

for help instead of the prof :”’))

Table 5.7 presents the top 15 features from unigrams (lowercase) and LIWC (uppercase) and the top 10 features from POS unigrams and bigrams correlated with bragging tweets grouped in six types.

We observe that **Achievement** statements usually involve verbs that are in past tense or indicate a result (e.g. *FOCUSPAST*, *finished*, *beat*). A POS pattern common in **Achievement** statements is a cardinal number followed by nouns in plural (i.e. *CD_NNS*), similar to its unigram and LIWC features (e.g. *NUMBER*, *3*, *5*) (e.g. *"I made a total of 5 dollars from online surveys wooo"*).

It is worth noting that one of the prevalent LIWC features for **Action** is *FOCUSFUTURE*. This is because the user may brag about a planned action (e.g. *"@USER You know what? I'm going to make some PizzaRolls Brag"*).

Most of the top words in **Feeling** express emotion or sensitivity (e.g. *happy*, *blessed*), which is consistent with the top POS feature, *RB_JJ* (e.g. *absolutely chuffed*, *so happy*).

In **Trait** category, words are mostly pronouns (e.g. *I*, *PRP*, *PRP_VBP*) and verbs (e.g. *VBP*, *VBP_JJ*).

Words that appear frequently in **Possession** category are actions related to purchase (e.g. *own*, *buy*) and nouns related to a tangible object (e.g. *car*, *bedroom*). In addition, users usually show off the value of their possessions using statements that involve currency signs (e.g. *$*) or currency signs followed by a number (e.g. *$_CD*) (e.g. *"I just signed a new three-year contract and I'll be getting 235 anytime minutes per month. Plus, the company is going to throw in a phone for just $49 per month. I'll bet you can't beat that deal!"*).

Finally, top words in **Affiliation** category involve positive feelings towards belonging to a group (e.g. *proud*, *amazing*) and nouns related to it (e.g. *FAMILY*, *team*).

5.5.2 Bragging and Post Popularity

We also analyze the association between bragging posts and the number of favorites/retweets they receive from other users. Similar to the previous linguistic feature analysis, we use univariate Pearson correlation to compute the correlations between the log-scaled favorites/retweets number of each tweet and its label (i.e. bragging or non-bragging) by controlling the numbers of followers and friends of the user who posted the tweet. Our results show that the number of favorites is positively correlated with bragging (see Figure 5.2) while there is no correlation between bragging and the number of retweets.

We further explore the popularity of different bragging types. We randomly analyze a

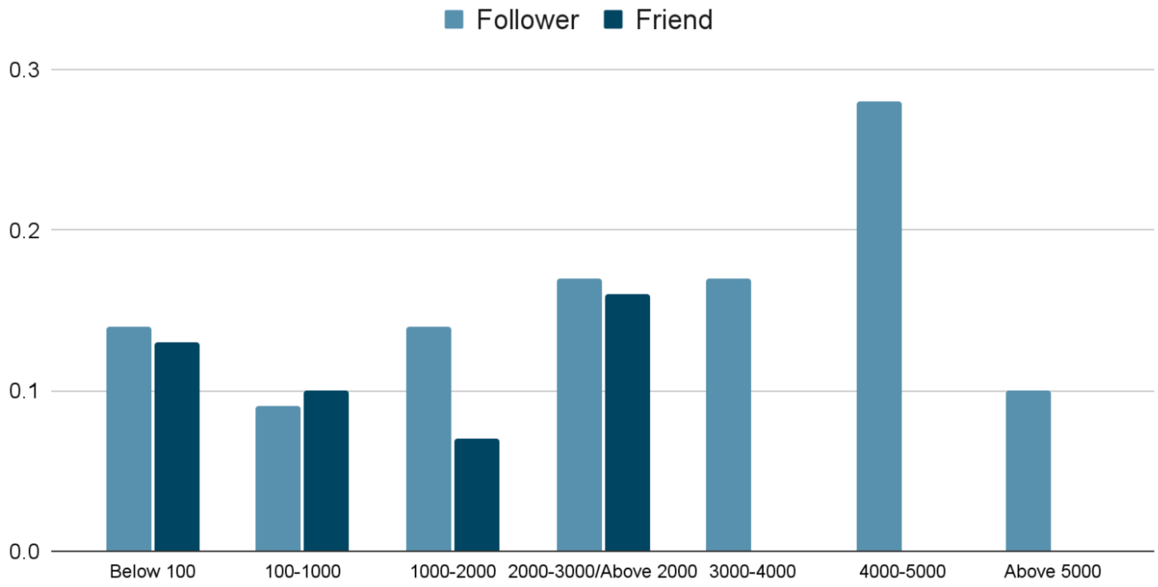


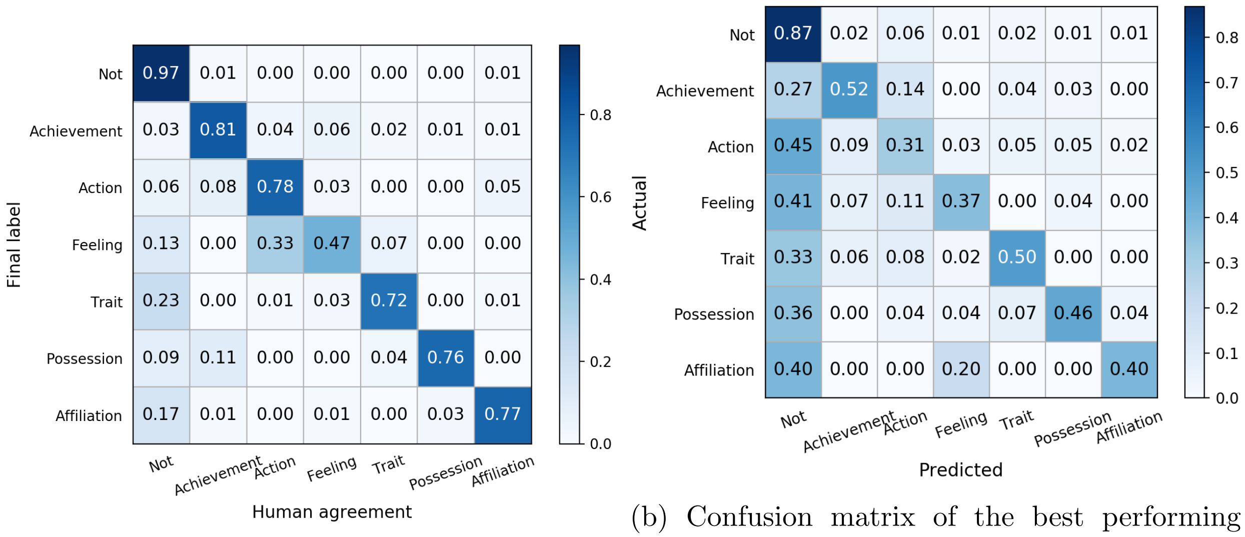
Figure 5.2: Pearson correlation between Twitter favorite number and bragging by controlling the number of followers and friends. All correlations are significant at $p < .01$, two-tailed t-test.

Class	**Mean**	**Median**
Achievement	3.06	3.00
Action	0.91	0
Feeling	0.50	0
Trait	2.38	2.00
Possession	2.00	0.50
Affiliation	5.50	2.00

Table 5.8: Mean and median Twitter favorites across bragging classes on a sample set of the data.

set of 443 tweets containing 56 bragging statements, where the follower and friend numbers of users are within a similar range: from 100 to 500 followers and from 500 to 1000 friends ($r = 0.19$, $p < .01$). We compute the mean and median Twitter favorites across the six bragging classes (see Table 5.8).

We observe that bragging statements about **Affiliation** such as family members or sports teams are more likely to receive a considerable amount of favorites with a mean of 5.5.



(a) Confusion matrix of annotator agreement on seven bragging categories.

(b) Confusion matrix of the best performing model on multi-class bragging classification, i.e. BERTweet-Clusters.

Figure 5.3: Confusion matrix on seven bragging categories.

For example, 14 users favorite the tweet *"This maybe is a little, but I'm SO proud of my research group. We represent so many different personality types, cultures, ways of thinking, etc, and every single member of my lab (all 21 of them)".* We speculate this is because praising the group that one belongs to instead of oneself as a bragging strategy enables users to be perceived as more likeable, especially by audiences who happen to be in the same group, which has been observed by Ren and Guo (2020). Furthermore, bragging about **Achievement** is generally marked as favorite by other users with a median of 3, where bigger achievements in the content such as job offers may receive more favorites (e.g. tweet *"Scored 80% on my thesis. Rather proud of that given the circumstances: new baby; pandemic; late topic change due to lockdown; minimal uni support because of furloughs; and an international move."* was marked as favorite 15 times).

5.5.3 Class Confusion Analysis

Figure 5.3a presents the confusion matrix of human agreement on seven classes normalized over the actual values (rows). We observe that **Non-bragging** (97%), **Achievement** (81%) and **Action** (78%) have high agreements, consistent with the class frequency. **Affiliation** (77%), **Possession** (76%) and **Trait** (72%) have comparable percentages as these are easily associated with a bragging target or group (e.g. family members, tangible objects or personalities). The **Feeling** category has the lowest percentage (47%) mostly caused by misclassification as the **Action** category (33%). This is due to the fact that both types are not associated with a concrete outcome by definition. Also, feelings are usually linked to an action. Thus, it makes the boundary between bragging about the action or the feeling associated with the action more challenging to interpret. The next most frequent confusion is between **Possession** and **Achievement**, which usually arises when a tangible possession is involved and the annotators disagree if the author was bragging about the actual possession or the action that leads to the author obtaining that possession (e.g. *"@USER I just got some stealth 300 easily the best headset I've ever had going from astro to turtle beach was a night and day difference"*).

Figure 5.3b presents the confusion matrix between bragging type labels and predictions by the best performing model, **BERTweet-Clusters**, on the multi-class classification task. First, we observe that the model is more likely to misclassify other classes as the dominant class, **Non-bragging**. Secondly, the most unambiguous classes are **Non-bragging** (87%) and **Achievement** (52%), which are in line with human agreement. Also, the model is good at identifying **Trait** (50%) and **Possession** (46%) due to the particular bragging targets (e.g. personalities, skills or tangible objects). Furthermore, we notice that the percentages of **Action** (31%) and **Feeling** (37%) are low. We speculate this is because they share more similarities with other classes (e.g. involving actions). This might also explain the high percentage of misclassified data points between **Action** and **Achievement**, **Feeling** and **Action**. Lastly, the model often confuses **Affiliation** with **Feeling** likely because the terms that express positive feelings (e.g. *"proud"*, 💙) also appear frequently in **Affiliation** (see Table 5.6).

5.5.4 Error Analysis

Finally, we perform an error analysis to examine the behavior and limitations of our best performing model (i.e. **BERTweet-LIWC** for binary classification and **BERTweet-Clusters** for multi-class classification) and identify pathways to improve the task modeling.

We first start with the binary bragging classification. We observe that **non-bragging** tweets containing positive sentiment are easy to be misclassified as bragging and even if such tweets involve something valued positively by authors, the purpose is usually to express recommendation, compliment or appreciation to others:

> T1: *"@USER paid for my* **new bottle of vodka** *&* **I Love Her with all my heart** ❤️*"*

Another frequent error happens when **non-bragging** tweets contain popular bragging targets such as achievement-oriented (e.g. *weight loss*, *marathon*) or possession-oriented (e.g. *car*, *electronics*):

> T2: *"4 spaces left on my budget* **weight loss** *program. £5 a week!???"*

Bragging often involves contextual understanding that goes beyond word use and requires a deep understanding of the context to determine the label. For example, common terms such as *first*, *finally*, *just* often appear in both **non-bragging** (T3) and **bragging** (T4) tweets:

> T3: "**just cleaned** *my cats' toilets"*

> T4: *"It happened again! I* **just completed** *30 minutes of meditation with @USER.* **Just** *sitting and resting in presence."*

Models also fail to detect **bragging** mainly because it is indirect or there are no typical trigger terms, so they lean on pre-training to contextualize:

> T5: *"9 hr drives feel like nothing now lol"*

Some **bragging** statements use additional mitigation strategies, e.g. re-framing the bragging statement as irony, as a complaint or invoking praise from a third party:

> T6: *"I find it strange how I was always the weird one in school and irl but* ***online people think*** *im cool for some reason"*

Finally, we highlight some representative examples of model confusion between bragging types. One example is when users' actions lead or do not lead to a concrete result. In this example the model predicted **Action**, but the actual label is **Achievement**:

T7: *"not to appropriate the gang escapes culture but me n my parents just did an escape room n actually got out?"*

Another example is an **Action** misclassified as **Possession**. This usually happens when a common phrase indicative of a certain type of bragging (e.g. *a new dish*) is invoked as part of an action:

T8: *"I had* ***a new dish*** *'egusi' it's so damn good! Love Nigerian food!"*

Other errors occur when multiple types of bragging are present (e.g. Feeling and Action) but the label expresses the more salient type, such as the **Feeling** highlighted in this example:

T9: *"Literally* ***had the best time*** *with the girls last night, don't think I've drank that much in my life?"*

5.6 Summary

In this chapter, we presented the first computational study to analyze and model bragging as a speech act along with its types in social media. We introduced a publicly available annotated data set in English collected from Twitter. We experimented using transformer models and incorporating linguistic information on bragging and bragging type prediction. The results showed that adding LIWC and topical features is beneficial to binary bragging identification and bragging type classification respectively. Finally, we presented an extensive analysis of features related to bragging statements, an analysis of associations between bragging posts and their popularity and an error analysis of model predictive behavior.

The best performing model in the bragging identification task (i.e. BERTweet-LIWC) is also used in Chapter 6 to analyze the relationship between bragging behavior online and user socio-demographic traits.

Chapter 6

Examining the Relationship between Bragging and Socio-demographic Factors

We have previously introduced the task of bragging identification and bragging type classification in Chapter 5. Bragging as a self-presentation strategy has benefits for the user when employed online such as gaining admiration, respect and attention from other users. Moreover, social media platforms build functionalities to reward and promote positive statements, e.g. by allowing users to follow each other and like each others' posts. Thus, many users pay great attention to building their online social image (Goffman et al., 1978; Baumeister, 1982; Leary and Kowalski, 1990) which makes self-promotion pervasive on social media (Dayter, 2018; Matley, 2020; Ren and Guo, 2020).

Large scale computational studies of social media content indicate online language is highly predictive of user socio-demographic traits such as age (Rao et al., 2010), gender (Burger et al., 2011), personality traits (Plank and Hovy, 2015) or education levels (Sujay et al., 2018). However, only limited work has been conducted in (computational) sociolinguistics to analyze the association between these social factors and self-promotion behavior (see Section 2.6). In addition, all these studies only focus on gender differences and do not examine other sociolinguistic factors that may have an influence on bragging (Altenburger et al., 2017; Wang et al., 2021).

In this chapter, we use a classifier (presented in Section 5.2.2) to estimate the prevalence

Label	Training set	Dev/Test set	All
Bragging	544 (16.09%)	237 (7.15%)	781 (11.66%)
Non-bragging	2838 (83.91%)	3077 (92.85%)	5915 (88.34%)
Total	3382	3314	6696

Table 6.1: Bragging data set statistics.

of bragging in a controlled and longitudinal data set that includes over 1 million English tweets posted by a group of 2,685 Twitter users in the U.S. over ten years. Then, we study the relationship between user socio-demographics and bragging. We also conduct an extensive linguistic analysis to unveil specific bragging themes and expressions associated with user traits and temporal factors.

It is important to clarify that this work is exploratory and indicative rather than definitive. The findings presented in this work are based on a small set of users in the U.S. and their posts during a certain time period. These results provide initial insights and trends but may not represent conclusive or absolute results. The study aims to offer a preliminary understanding of the topic and to pave the way for further research and investigation.

The work presented in this chapter has been submitted to EPJ Data Science.

6.1 Measuring Bragging Prevalence

6.1.1 Predictive Model

We use the best performing predictive model on identifying whether a tweet contains bragging or not (i.e. **BERTweet-LIWC**) introduced in Chapter 5. The BERTweet-LIWC model encodes texts using BERTweet (Nguyen et al., 2020) and combines them with Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2001) features. The feature combination is performed through a fusion mechanism called multi-modal adaption gate, which was originally introduced by Rahman et al. (2020). The joint representations of texts and LIWC features are finally sent to an output binary classification layer (see Section 5.2.2).

6.1.2 Bragging Prevalence Metrics

We measure the prevalence of bragging for (1) all users in a given time period; and (2) single user individually.

All Users To measure the bragging prevalence for all users in our analysis data (see Section 6.2), we first obtain the distribution of bragging tweets and total tweets posted by all users over each year and month (denoted as $P = \{p_1, ..., p_n\}$ and $Q = \{q_1, ..., q_n\}$ respectively, where n is the number of months in the data set). We calculate a distribution of average bragging percentage across all users for each month $A = \{a_1, ..., a_n\}$ such that:

$$A = \frac{P}{Q} \tag{6.1}$$

Individual User To compute a bragging score of an individual user, we first obtain the distribution of bragging tweets and total tweets over each year and month for each user, which are denoted as $B_u = \{b_{u1}, ..., b_{un}\}$ and $T_u = \{t_{u1}, ..., t_{un}\}$ respectively. We obtain a time-normalized bragging distribution $D_u = \{d_{u1}, ..., d_{un}\}$ for each user over months by dividing each data point from the user distribution by the fraction of bragging tweets for all users in the time window:

$$D_u = \frac{B_u}{T_u * A} \tag{6.2}$$

Finally, we obtain the bragging percent l for each user by averaging the normalized bragging distribution D_u:

$$l = \frac{\sum_{i=1}^{n} d_{ui}}{n} \tag{6.3}$$

The normalized bragging percent practically compares the bragging tendency of a single user to that of the population average in the same time range. We normalize with time to account for possible temporal shifts in bragging prevalence over time (see Figure 6.1).

6.2 Analysis Data

To analyze the social factors of bragging, we need a large set of Twitter users associated with socio-demographic characteristics. We combine three data sets that contain such information in the following papers as provided by the authors on request: the first data set (developed by Preotiuc-Pietro et al. 2016) contains 863 users, the second data set (developed by Guntuku et al. 2017) contains 4,568 users and the third data set (developed by Jaidka et al. 2020) contains 938 users. All users are mapped to their self-reported genders and ages. The users in the second data set self-report their education degrees and annual incomes. In total, our combined data set contains 6,369 users (all from the U.S.).

6.2.1 Data Pre-Processing

Subsequently, we collect all historical tweets from these users resulting in more than 9.7 million tweets. We pre-process the data as follows.

Filtering First, we filter out non-English content using the language code provided by Twitter. Second, we exclude replies and retweets by setting related parameters in user timeline querying. We perform this step, as retweets are not original posts created by the user and replies were not used in the annotated training data, hence the model was not trained on this data. We also exclude extremely short tweets (i.e. containing less than three tokens) as these are very likely not to contain user bragging or non-bragging statements. Then, we remove the duplicate tweets by using the first five content tokens (excluding numbers, usernames and hashtags) because duplicate tweets are likely to be automated and they are not the original content created by users.

All tweets that are also automatically generated from third parties are filtered out using source labels that indicate original authorship such as “Twitter Web Client“, “Twitter for Android” or “Twitter for iPhone”. We found that tweets with source labels such as “The Sims 4 Game” and “Paradise Island 2” (e.g. *“I played the Sandy Caps mini game in Paradise Island 2, and my score was: 68 #ParadiseIsland2 #GameInsight”*) are likely to be generated automatically and would negatively impact our analysis. Finally, we remove all users that have posted fewer than 20 tweets. This is because computing an average bragging ratio

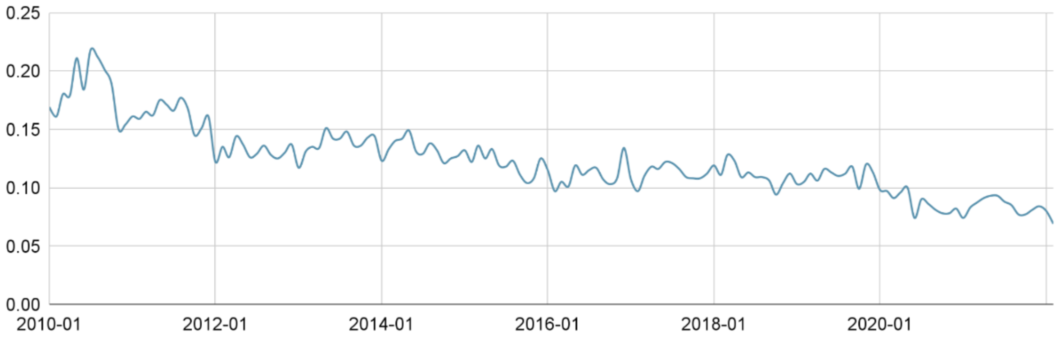


Figure 6.1: Bragging percentage by year and month.

across a very small number of tweets would be unreliable.[1]

In total, our analysis data contains 2,685 users with 1,031,276 tweets, with each user having 384 posts on average.

Text Processing We pre-process the collected tweets by lower-casing and tokenizing using TweetTokenizer from NLTK Toolkit.[2] We also replace all URLs and username mentions with a single word token <URL> and <USER> respectively.

6.2.2 Computing Bragging Ratio

We use the bragging predictive model to identify the category (bragging/non-bragging) of all 1,031,276 individual tweets in our analysis data set. To investigate the performance of the model on the analysis data, we manually evaluated a batch of 100 tweets across different users and years and found that the model achieved 78.55 macro F1, which was higher than the performance on the annotated test data set (72.42 macro F1).

Finally, we compute the normalized bragging ratio for each user (see Section 6.1.2 for the bragging metric definition). The mean normalized bragging ratio is 0.0020 and the median bragging ratio is 0.0013 for all users.

Figure 6.1 shows the bragging percentage over time across the ten years of the data set. This shows that overall the bragging percentage decreases with time and it highlights the need for temporal normalization of the bragging ratio.

[1] Early experimentation showed that a threshold of 20 tweets or larger leads to consistent correlations.
[2] https://www.nltk.org/api/nltk.tokenize.html

Trait	**Mean**	**Median**
Socio-demographic		
Age	35.99	33
Popularity		
No. Followers	22,951.06	186
No. Friends/No. Followers	2.62	1.58
No. Listings	47.72	2.0

Table 6.2: User socio-demographic statistics.

6.2.3 User Socio-demographic Traits

We examine the relationship between bragging behavior online and the following user socio-demographic traits:

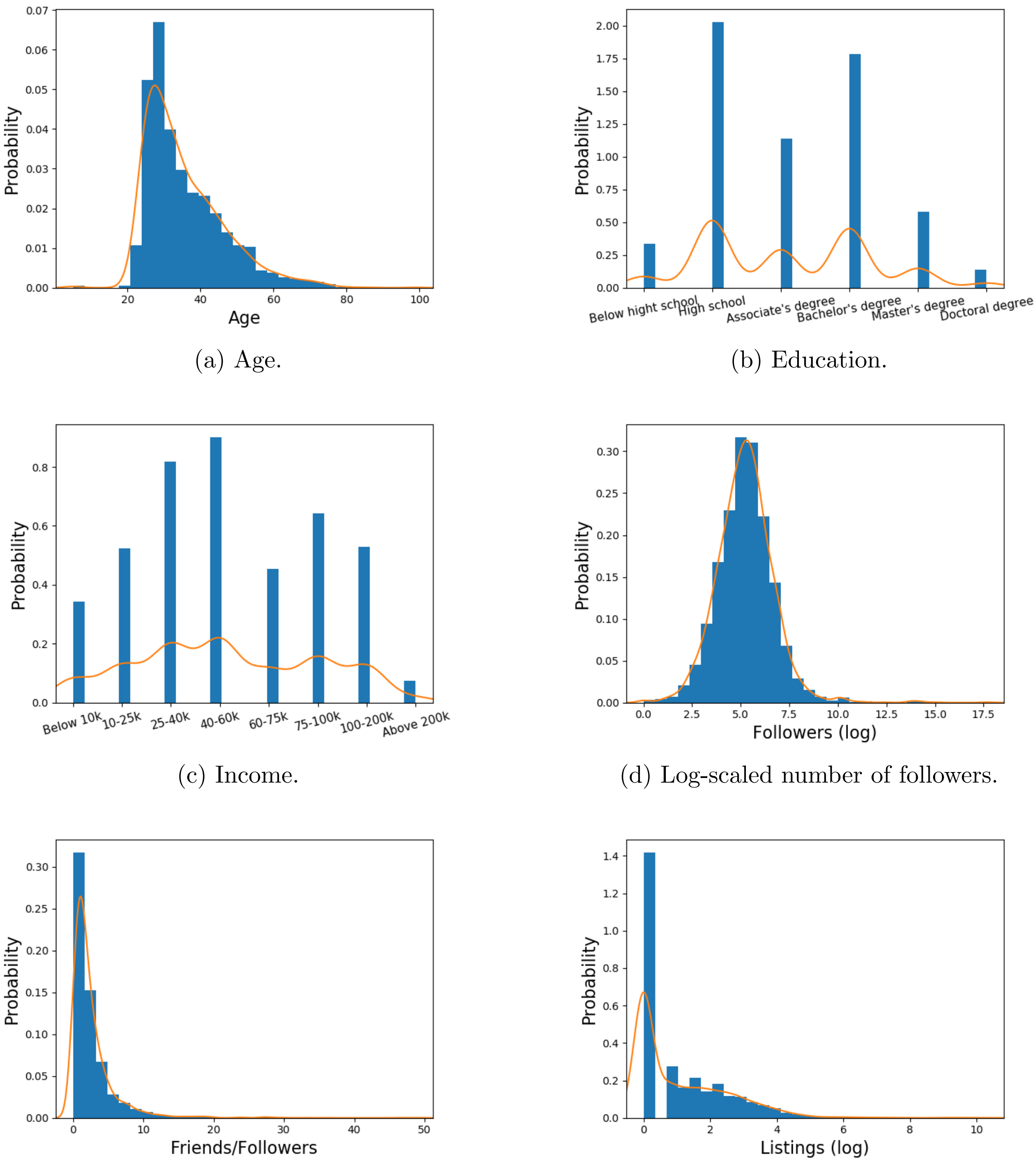
- *Gender.* There are 33.81% self-identified males labeled as 0 and 66.19% self-identified females labeled as 1. Self-identified non-binary users represented a very small number of the total participants and hence were removed from the analysis.

- *Age.* Reported as the year of birth. The age used throughout the paper is age as of the end of 2021.

- *Education.* It contains six categories: (1) users who have not completed high school; (2) high school or equivalent; (3) associate's degree or equivalent; (4) bachelor's degree or equivalent; (5) master's degree or equivalent; and (6) doctoral degree or equivalent.

- *Income.* The annual yearly income of a user in U.S. dollars is divided into eight categories: (1) below 10K; (2) 10-25K; (3) 25-40K; (4) 40-60K (5) 60-75K; (6) 75-100K; (7) 100-200K; and (8) above 200K.

- *Popularity.* Popularity reflects other people's interest in users' accounts or posts, which can be quantified by the number of followers, the ratio of friends and followers or the number of times a user was listed. Note that, for computing correlations between popularity metrics and bragging, we scale the number to make their distribution closer to a Gaussian. We collect all user information up until the end of 2021 to be consistent with the user age.



(a) Age.

(b) Education.

(c) Income.

(d) Log-scaled number of followers.

(e) Ratio of friends and followers.

(f) Log-scaled number of listings.

Figure 6.2: Histograms of user socio-demographic traits.

Table 6.2 shows summary statistics of the socio-demographic traits in the analysis data. The trait distributions in the data set are presented in Figure 6.2.

Trait	**Correlation**	p_{unc}	p_{corr}
Socio-demographic			
Gender (Female-1, Male-0)	0.10	<.001	<.001
Age	-0.16	<.001	<.001
Education	0.14	<.001	<.001
Income	0.07	<.003	<.002
Popularity			
No. Followers	0.12	<.001	<.001
No. Friends/No. Followers	-0.10	<.001	<.001
No. Listings	0.09	<.001	<.001

Table 6.3: Pearson correlations and p-values between user-level traits and their bragging metric. p_{unc} refers to uncorrected p-values and p_{corr} refers to corrected p-values using the Bonferroni correction (a method used to adjust p-values in statistical hypothesis testing to account for multiple comparisons).

6.3 Socio-demographic Factors and Bragging Prevalence

We study the relationship between bragging rate and user traits. We perform a correlation analysis by computing the Pearson correlation coefficient between user traits and the user-level bragging metric described in Equation 6.2 and Equation 6.3. The results are summarized in Table 6.3.

We first analyze gender and age. According to our analysis, gender and age are strongly associated with the bragging percentage ($p < .001$). Note that correlations around 0.1 with such a large sample size (N=2,685) are highly significant and in terms of magnitude in line with correlations between other well known linguistic variables and traits (Carey et al., 2015; Holgate et al., 2018).

By considering multiple intersecting identities, such as gender, age, income and education, the analysis can reveal how different combinations of these traits may influence bragging behavior differently, which leads to a more comprehensive understanding of the findings. The intersectional identities are unique and their combined effects can be greater than the sum of their individual categories (Preddie and Biernat, 2021). However, due to limited data availability, it is challenging to explore all possible intersectional categories. By controlling for certain variables, we attempt to reduce the potential confounding effect of intersectional categories and focus on understanding the individual effects of specific traits on bragging behavior. Thus, we examine the rest of the demographic traits by controlling for gender and age using partial correlation, where education level and annual income are represented in the ordinal scale described in the data set section (see Section 6.2.3).

The main findings are:

- Gender is significantly correlated with bragging in the sense that females brag more than men ($r = 0.10$, $p_{unc} < .001$, $p_{corr} < .001$). This is consistent with the findings of previous studies related to self-presentation (Sheldon, 2013; Wang et al., 2021) and it can be explained by the fact that women show more interest in developing friendships online (Holmes, 2013), which can be accomplished by positive self-presentation.

- There is a significant association between age and bragging, with younger users bragging more than older ones ($r = -0.16$, $p_{unc} < .001$, $p_{corr} < .001$). This might result from younger people's desire for increasing their status among peers and peer approval (MacIsaac et al., 2018). Social comparison was found to occur more frequently in younger age groups than in older ones (McAndrew and Jeong, 2012), which explains why younger users tend to create positive self-presentations online (Yau and Reich, 2019).

- Users with higher education levels tend to brag more ($r = 0.14$, $p_{unc} < .001$, $p_{corr} < .001$). This might be explained by the fact that users with a higher educational level tend to express more joy which could include self-disclosure statements (Volkova and Bachrach, 2015).

- Users with higher income significantly brag more frequently online ($r = 0.07$, $p_{unc} < .003$, $p_{corr} < .002$) than users with lower incomes. Users with higher income were found to be more likely to produce positive tweets (Volkova and Bachrach, 2015). Previous work also suggested that rich people are characterized by a self-focused and narcissistic personality (Leckelt et al., 2019), which leads to producing more that is related to self-promotion in social media (Buffardi and Campbell, 2008; Moon et al., 2016).

- Higher income and education are positively correlated with higher age (income: $r = 0.16$, $p_{unc} < .001$, $p_{corr} < .001$; education: $r = 0.30$, $p_{unc} < .001$, $p_{corr} < .001$). Higher age however has an inverse relationship to bragging than income and education. This highlights a divergence along these traits, where users who are either highly educated or young are likely to brag more.

- Users with a larger number of followers brag significantly more ($r = 0.12$, $p_{unc} < .001$, $p_{corr} < .001$). Similarly, the same trend applies to the lower ratio of friends and followers ($r = -0.10$, $p_{unc} < .001$, $p_{corr} < .001$) and the larger number of listings ($r = 0.09$, $p_{unc} < .001$, $p_{corr} < .001$). It is possible that users with many followers (e.g. micro- or macro-influencers) tend to interact with or try to maintain or obtain followers (Guo and Ren, 2020) by establishing a positive social image through bragging. This can be explained by the fact that users are more willing to share content with positive sentiment with people that share a weak relationship (e.g. online followers) than with actual real-life friends (Bak et al., 2012).

6.4 Influence of Socio-demographic Factors on Bragging

We further explore the relationship between language use and bragging behavior across different socio-demographic characteristics. In addition, we also highlight the use of bragging words at different days of the week and times of the day to shed further light on this phenomenon. We use a unigram (i.e. token) feature analysis to identify words and themes associated with bragging by computing the correlations between the distribution of each unigram across posts and the label of the post (i.e. bragging or not bragging). Then, we use univariate Pearson correlation to rank the unigrams similar to Schwartz et al. (2013).

6.4.1 User-Level Analysis

We first examine the individual differences in bragging on Twitter by gender, age, education level and income.

Gender Table 6.4 (left) shows the top 25 unigrams correlated with gender. We observe that males mostly brag about their partners (e.g. *wife*), but also mention other users (e.g. <*user*>). Popular bragging topics for males are entertainment achievements such as games

Male vs. Female				**Born after 1988 vs. before 1988**			
Unigram	r	Unigram	r	Unigram	r	Unigram	r
<url>	.073	my	.083	i	.099	<url>	.141
<user>	.071	so	.063	my	.069	<user>	.079
game	.061	😍	.060	😍	.069	!	.077
team	.047	:)	.059	me	.057	join	.036
games	.040	:p	.058	so	.050	#livepd	.033
#dnd	.036	❤	.048	😭	.050	local	.032
#league	.036	hair	.048	m	.049	kids	.031
wife	.034	mom	.043	😂	.047	our	.031
tournament	.034	☺️	.042	😊	.041	we	.031
podcast	.033	🙃	.041	class	.040	wife	.031
win	.033	bed	.041	exam	.039	daughter	.030
#livepd	.033	love	.040	college	.038	via	.029
football	.033	😂	.040	semester	.038	w	.029
#twitch	.033	😊	.040	lol	.037	play	.028
fantasy	.033	me	.039	myself	.036	#dnd	.028
great	.032	😳	.037	life	.036	pe	.028
stream	.032	♀	.036	☺️	.036	inboxdollars	.028
aw_prints	.031	💁	.035	ve	.035	app	.028
inboxdollars	.031	😁	.034	best	.035	challenge	.027
championship	.030	😭	.034	🙃	.035	covet	.026
playing	.030	🙈	.033	🙄	.035	show	.026
teams	.030	because	.033	💁	.035	awesome	.025
congratulations	.029	i	.033	😅	.035	#positive	.025
app	.029	happy	.033	like	.034	photo	.025
bout	.029	boyfriend	.033	mom	.034	#i_am	.024

Table 6.4: Unigram feature correlations with bragging between gender (left) and age (right), sorted by Pearson correlation (r). All correlations are significant at $p < .001$, two-tailed t-test.

(e.g. *#dnd*, *#league*, *#twitch*, *stream*) and sports (e.g. *tournament*, *win*, *football*, *championship*, *teams*). For example,

> T1: *"I'm so genuinely happy I witnessed that. #NationalChampionship"*.

On the other hand, females prefer to brag by using first person pronouns (e.g. *my*, *me*, *i*), bragging about personal traits (e.g. *hair*), feelings (e.g. *love*, *happy*) and their partners (e.g. *boyfriend*). For example,

> T2: *"HOORAY!! I finally got a haircut appointment!! It'll be 64 days since my last cut (I normally go every 5 weeks so this'll be close to double my normal tim, and yeah, it feels like I have about twice as much hair to cut!!)"*.

Studies have shown that women spend more energy than men in presenting themselves for impression management (McAndrew and Jeong, 2012). Furthermore, bragging by females usually contains female-related terms (e.g. ♀) or positive emoticons (e.g. :), :p) and emojis (e.g. 😍, ❤️) to strengthen the meaning of their posts. This corroborates the findings that positive emojis are used more frequently in positive contexts (Derks et al., 2008). Results are also consistent with the observation that women communicate using more emotional exchanges (Gefen and Ridings, 2005). They also use emojis more often than men (Chen et al., 2017; Prada et al., 2018).

To further investigate how women use emojis while bragging, we randomly choose 100 tweets with emojis generated by women which are classified as bragging. In addition to the wrongly predicted tweets (23%), tweets with one emoji only are almost twice as many as tweets with multiple emojis (51% vs. 26%). The most popular type is related to positive ones (43%) including smiling faces and hearts, which also can be observed in Table 6.4 (left). We notice that smiling faces are used to express speakers' happiness and excitement, e.g.

> T3: *"I am finally understanding accounting! I feel so smart! YAYY!* 😊 😄 *<URL>"*

while bragging with heart emojis usually involves other people or objects, e.g.

> T4: *"thomas is watching camp rock with me, it's safe to say he is the best bf* ❤️ ❤️*"*.

Other emoji faces (29%) such as 😂, 🙃, 😩 are mostly used to convey a joke or irony or to soften the magnitude of bragging, e.g.

T5: *“Cash in some flight miles... im there! 😂”*.

Emojis with gesture (10%) such as 🙌, 👌 are used to express pride for a victory, e.g.

T6: *“The parking gods are with me today! YESSS! 🙌 🙌 🙌”*

while other emojis (11%) represent accomplishments, e.g.

T7: *“Take it out on the gym ✅”*

or specific topics, e.g.

T8: *“Took the kid for a little football 🏈 training... he is growing like a weed!! <URL>”*.

Age In terms of age, we split the users into two age groups (born before 1988 and after 1988) using the population median. Table 6.4 (right) shows the top 25 unigram features correlated with age. We notice that younger users (born after 1988) brag more about themselves (e.g. *i*, *my*, *me*, *myself*) and school life (e.g. *class*, *exam*, *college*, *semester*). For example,

T9: *“Also somehow got a 90 on my soils exam... how...”*.

Also, they use more emojis in their bragging posts, which is consistent with the fact that in general younger users tend to use more emojis than older ones do in social media (Prada et al., 2018). The group of users above median age brags more about collective activities (e.g. *our*, *we*) and their affiliation such as family members (e.g. *kids*, *wife*, *daughter*), which suggests older people are more family-focused and engage more with family activities (McAndrew and Jeong, 2012). For example,

T10: *“My daughter is hands down the coolest person I know!”*.

Education & Income Table 6.5 presents the top 15 unigrams in bragging posts correlated with higher and lower education (left) and higher and lower income (right). For education, we observe that people with higher levels of education use a smaller number of emojis. They brag more about their jobs (e.g. *student*, *conference*) and activities involving food (e.g. *beer*, *delicious*). For example,

> T11: *"Glad I've been working on my Adobe Suite skills this year. I'm going to be making a very special obituary poster for my uncle's celebration of life to highlight testimonials from friends and family* 💜*".*

Furthermore, people with higher education like to mention others (e.g. <*user*>, *our*, *their*) in their bragging statements. This is in contrast with lower educated users who focus on their personal traits, possessions or activities (e.g. *i*, *my*, *look*, *got*). For example,

> T12: *"I look cute as hell in this hoodie".*

Similar income and education levels lead to a similar language (e.g. <*user*>, *students*, *congratulations* in higher education and higher income, *my*, *lol*, *baby*, *boyfriend* in lower education and lower income). However, bragging expressions from higher income users are more related to entertainment events such as *#music*, *golf* compared with higher education. For example,

> T13: *"Great golf lesson with <USER> at the PAGA! Already see the results. Now-practice to see them more often. #GonnaBreak80".*

6.4.2 Bragging Language and Time

Finally, we complement the user-level language analysis by examining the differences between ways of expressing bragging across different days of the week and times of the day.

Day of Week Table 6.6 presents the top 20 unigram features correlated with bragging on weekdays versus weekends. For this analysis, we normalize the creation time of each post to the local time by using the timezone difference which was inferred from the zip code that users have provided.

Higher Education vs. Lower Education				**Higher Income vs. Lower Income**			
Unigram	r	Unigram	r	Unigram	r	Unigram	r
<user>	.062	i	.074	<user>	.061	:)	.043
students	.040	lol	.059	<url>	.051	my	.030
<url>	.039	my	.055	aw_prints	.032	lol	.030
our	.030	😍	.043	congratulations	.030	got	.026
congratulations	.029	m	.038	#iteachk	.029	#livepd	.023
pe	.025	👌	.035	pe	.029	work	.023
ss	.025	look	.033	#piano	.029	cory	.022
season	.025	😂	.032	#happyclassrooms	.027	covet	.022
zak	.025	got	.032	students	.026	renee	.021
conference	.024	😁	.031	#music	.024	makes	.021
beer	.024	☺️	.030	team	.024	boyfriend	.021
#piano	.023	#10billionwives	.028	golf	.024	yay	.021
their	.023	😊	.027	win	.024	m	.020
delicious	.023	baby	.027	💙	.023	then	.019
student	.021	😭	.026	bet	.023	kiss	.019

Table 6.5: Unigram feature correlations with bragging between higher and lower education level (left) and higher and lower annual income (right), sorted by Pearson correlation (r). All correlations are significant at $p < .001$, two-tailed t-test.

To demonstrate the face validity of the analysis, we first observe that bragging statements from both weekdays and weekends involve words related to time (e.g. *thursday*, *sunday*, *night*, *weekend*, *afternoon*). Secondly, we observe that users mostly brag about their school life or work on weekdays (e.g. *class*, *professor*, *interview*, *office*, *internship*). For example,

> T14: *"My professor for accounting saw what I carry all day and said I must have great upper body strength lol".*

Another popular bragging topic on weekdays is about going to the *gym*. Bragging on weekends usually focuses on certain entertainment, recreation and worship activities (e.g. *church*, *watching*, *bar*, *football*, *drinking*, *party*). For example,

> T15: *"Yay, I'm picking up my cute new glasses tomorrow. Now I can rock them*

Weekday vs. Weekend			
Unigram	r	Unigram	r
class	.050	sunday	.034
professor	.028	night	.029
semester	.028	friday	.028
job	.026	church	.028
campus	.024	<user>	.026
classes	.023	#livepd	.025
#tbt	.023	weekend	.025
thursday	.021	we	.024
interview	.021	state	.024
exam	.021	watching	.020
office	.020	bar	.020
bio	.020	game	.020
monday	.020	won	.019
killed	.019	football	.019
grow	.019	drinking	.019
vote	.019	rewards	.018
internship	.018	afternoon	.018
ago	.018	party	.018
teeth	.017	racing	.018
gym	.017	jam	.017

Table 6.6: Unigram feature correlations with bragging between weekday and weekend, sorted by Pearson correlation (r). All correlations are significant at $p < .001$, two-tailed t-test.

at Mom's party tomorrow".

In addition, this could also involve activities that are done as part of a group, as exemplified by the first person plural pronoun *we*.

06:00-09:00		09:00-12:00		12:00-15:00		15:00-18:00		18:00-21:00		21:00-00:00		00:00-03:00		03:00-6:00	
Unigram	r	Unigram	r	Unigram	r	Unigram	r	Unigram	r	Unigram	r	Unigram	r	Unigram	r
morning	.131	morning	.042	lunch	.053	dinner	.040	dinner	.042	tomorrow	.049	i	.055	sleep	.069
last	.055	coffee	.040	just	.029	<url>	.025	#piano	.031	tonight	.047	night	.044	awake	.069
breakfast	.055	class	.037	afternoon	.023	store	.022	#music	.026	bed	.040	sleep	.043	morning	.069
today	.051	last	.031	shopping	.020	pizza	.021	tonight	.024	i	.038	midnight	.041	early	.058
day	.048	today	.029	basically	.020	came	.021	beer	.022	night	.029	bed	.032	up	.057
up	.046	lunch	.028	outside	.020	bags	.020	tomorrow	.020	#livepd	.026	😪	.032	4am	.051
sleep	.044	yesterday	.028	shop	.020	fitness	.018	#musicmonday	.020	midnight	.026	2:30	.031	5:30	.047
coffee	.044	woke	.028	break	.020	mail	.018	pizza	.021	win	.025	drunk	.031	6:30	.042
school	.044	breakfast	.027	cleaning	.019	published	.018	#classicalmusic	.021	life	.025	tonight	.030	covet	.041
yesterday	.043	classes	.025	food	.019	rain	.017	wine	.021	absolutely	.023	friends	.030	singles	.039
work	.043	#protapes	.025	burger	.019	<user>	.017	watching	.021	friends	.022	1am	.029	4:30	.039
woke	.042	iced	.024	:d	.019	done	.017	correct	.020	boyfriend	.020	insomnia	.028	wide	.038
early	.041	into	.022	books	.018	recipe	.016	cookies	.020	alive	.019	m	.028	scored	.038
class	.039	office	.021	shirt	.018	secretly	.016	drinking	.020	legal	.019	best	.028	wake	.037
then	.036	day	.021	children	.016	dryer	.016	garden	.020	seen	.019	onlyfans	.028	diddy	.035

Table 6.7: Unigram feature correlations with bragging between different time periods in a day, sorted by Pearson correlation (r). All correlations are significant at $p < .001$, two-tailed t-test.

Time of Day Table 6.7 shows the unigrams associated with bragging between different times in a day. Similar to the day of the week analysis, our findings demonstrate face validity since the top correlated tokens are related to the time of bragging (e.g. *morning*, *yesterday*, *tonight*, *2:30*). Next, we observe that bragging about eating or sharing food such as *coffee*, *lunch*, *pizza*, *beer*, *cookies* is popular at all times except at late night. For example,

> T16: *"Just made some lamb burgers with homemade tzatziki sauce. Starting to feel confident about my cooking skills."*.

Also, bragging in the morning is usually about things that happened the previous day (e.g. *yesterday*, *sleep*) or study/work (e.g. *school*, *class*, *office*, *"Secured a B in my principles of marketing class this semester!!!"*) while bragging in the afternoon or evening involves a wide variety of recreational activities (e.g. *shopping*, *fitness*, *#music*, *watching*, *"Finally, newest member of planet fitness!"*). Finally, many users tend to brag about their upcoming activities in the evening (e.g. *tomorrow*, *morning*).

6.5 Limitations

The findings of our study focus on users based in the U.S. that regularly post on Twitter. Analyzing this specific set of Twitter users located in the U.S. offers valuable insights, but it also has inherent limitations in terms of representativeness. Firstly, the samples may not represent the entire global Twitter user population and findings may not apply to users from other countries or regions. Secondly, Twitter itself attracts a particular user base with specific preferences and interests. Analyzing only this subset of users may not capture the behavior and attitudes of users on other social media platforms or those who do not use Twitter at all. Finally, the use of an English-only dataset may introduce language bias, potentially leading to the underrepresentation of non-English speaking users.

Moreover, we elicited self-reported demographic traits of users a single time through a survey obtained in 2016, while the data we analyze spans from 2010 to 2022. We expect that this could impact education and income results, although the signal should still be indicative of actual education and income for the most part of the tweet post times. Users in the survey could self-identify as non-binary gender, but due to the small sample size of this category, we were only limited to studying binary gender and removed non-binary gender users. Finally, the user popularity metrics such as the number of followers are obtained at the end of 2021 and may have evolved since the tweets used to compute their bragging ratio were posted. However, the ranking in popularity across users is mostly stable across years, so this is unlikely to impact our correlation results.

6.6 Summary

In this chapter, we presented the first large-scale empirical study of the relationship between bragging and user socio-demographic factors in computational sociolinguistics. Our analysis of more than 1 million English Twitter posts from users in the U.S. showed that females, younger users and users with higher education, higher income and popularity tend to be more braggarts than other users. Finally, through feature analysis, we were able to identify the popular topics and expressions of bragging across different users at different times. These results serve as a starting point for future studies and should be interpreted as preliminary indications rather than definitive conclusions.

Chapter 7

Conclusions

This thesis presented a computational study of two common speech acts, complaining and bragging, in social media. This chapter summarises the tasks, findings and contributions presented throughout the thesis, discusses potential ethical implications and indicates possible directions for future work.

7.1 Summary of Thesis

Chapter 2 first presented background knowledge on speech acts in pragmatics. Following this, we focused on two common speech acts in daily communication, complaining and bragging, by describing their definitions, types and pragmatic strategies. We also presented previous work in the field of linguistics and psychology. Then, we introduced speech act detection as a text classification task including definitions and popular models. Finally, we reviewed tasks and approaches for modeling complaints and self-disclosure in NLP and identified their limitations.

Chapter 3 introduced the task of classifying complaint severity levels as well as a new annotated data set consisting of four severity levels. The data set has been made publicly available. We proposed a method to incorporate external linguistic features into transformer networks. The results showed that introducing emotional and topical information to the model is beneficial to complaint severity classification. We also analyzed the behavior of our models in predicting complaint severity levels. We found that models struggled with cases where texts were expressed in a subtle way or belonging to neighboring levels.

Chapter 4 compared a series of transformer-based models and models that use external linguistic knowledge and are pretrained on a large "noisy" data set in a complaint identification task. We also proposed two MTL models to jointly model complaint identification (main task) and complaint severity level classification (auxiliary task). The results showed that the severity level information in MTL settings helped improve the predictive performance of complaint identification and achieved state-of-the-art performance. We also identified the limitations of models through an error analysis, where it is easier for models to misclassify complaints with implicit and ironic expressions as non-complaints.

Chapter 5 introduced a new data set annotated with bragging and its types. The data set has been made publicly available. We evaluated transformer networks with different linguistic features on bragging identification and bragging type classification. We showed that adding LIWC and topical information resulted in higher predictive performance of binary bragging identification and bragging type classification respectively. We also presented an extensive analysis discovering linguistic patterns of bragging statements, the popularity of different bragging types and model predictive behaviors.

Chapter 6 presented a large scale empirical study on examining the relationship between online bragging and common user socio-demographic factors (i.e. gender, age, education, income and popularity). We used a bragging classifier to predict over 1 million English Twitter posts from 2,685 users in the U.S.. We measured the prevalence of bragging for all users and individual users by normalizing it with time. The results showed that females, younger users and users with higher education, higher income and popularity are more likely to brag on Twitter. We also performed a linguistic feature analysis revealing bragging language use across different socio-demographic characteristics and times.

7.2 Ethical Implications

The development and deployment of models to classify complaints and bragging in social media have several ethical implications that need to be carefully considered. Note that our work has received approval from Ethics Committee of our institution and official Twitter API. This section will discuss ethical considerations including the possible applications of such models, potential misuse and privacy concerns.

Possible Applications The primary aim of these classification models is to enhance online communication by providing users with valuable feedback on their posts. These classifiers have broad applications, including warning users of sounding excessively boastful or complaining too frequently. By identifying and flagging content that might come across as boastful or complaining, individuals can be encouraged to modify their online behavior to maintain more positive and respectful interactions on social media platforms. Moreover, beyond individual users, this technology enables companies to identify and address customer complaints more effectively and thus enhances customer experiences and maintains a positive brand image.

Possible Misuse and Mitigation Strategies As with any technological advancement, there is a risk of misusing these classification models by malicious actors or even individuals who might be unaware of the ethical considerations. For instance, people could exploit the models to falsely categorize someone's content as bragging or complaints, leading to harassment or defamation. To mitigate potential risks associated with these applications, continuous evaluation and transparent development are crucial. Regular assessments of the model's performance and impact can identify any unintended consequences and enable appropriate adjustments. Educating users about the potential risks of misuse is equally essential. Raising awareness about the limitations of these models can help users interpret the results more critically and avoid drawing hasty conclusions or making uninformed judgments about others. Moreover, providing use guidelines and ethical principles encourages responsible use to prevent misinterpretation and undue consequences for individuals.

User Privacy The work presented in Chapter 6 used sociodemographic data, raising concerns about the responsible handling of personal information when conducting population-level social media research. First, researchers must prioritize data anonymization and protection to ensure the privacy and confidentiality of users. Furthermore, stakeholders must comply with relevant data protection regulations, establish clear guidelines for collecting, storing, and retaining user data and mention if the data is shared with third parties. Additionally, the research focuses on investigating bragging behavior online at the population level instead of the individual. Given a large-scale data set, individual users are less likely to be specifically identified (Conway and O'Connor, 2016).

Involvement of Company The work presented in Chapter 5 and Chapter 6 involves collaboration with a research scientist from Bloomberg. In the initial stage of data annotation (i.e. testing round), Bloomberg provided funding to hire crowdsourced workers on Amazon Mechanical Turk (MTurk). However, due to the low inner agreement, we decided to annotate by ourselves without pay. Later, Bloomberg conducted an internal review to make sure the publications were in compliance with the company policies for research. Moreover, an internal reviewer at Bloomberg provided feedback on our paper (Jin et al., 2022) due to his/her interest. Apart from that, the research process and results are independent and all data and code is publicly available for research purpose.

Lastly, it is essential for readers and users to understand the limitations of these models and research. The findings should be viewed as exploratory results rather than conclusive assessments. Users should be encouraged to critically analyze the results and consider the context before drawing any conclusions. Meanwhile, researchers and companies should avoid making sweeping generalizations based solely on predictions. Instead, the findings should be considered as tools to aid decision-making and complement human judgment.

7.3 Future Directions

The research work in this thesis can be extended or generally used in other NLP applications. We mention some directions for further research:

- The method for injecting external linguistic information (e.g. LIWC, topic, emotion) into transformer-based models proposed in Chapter 3 can be applied in other classification tasks such as hate speech detection (Mozafari et al., 2019) and sarcasm detection (Srivastava et al., 2020).
- The multi-task learning approaches presented in Chapter 4 increased the performance of the primary task (complaint identification) by benefiting from the auxiliary task (complaint severity classification). These approaches can be extended to other classification tasks for jointly modeling two or more related tasks such as metaphor identification and emotion prediction (Dankers et al., 2019), stance detection and sentiment analysis (Upadhyaya et al., 2022).
- The task of identifying complaints presented in Chapter 4 can be extended to multimodal (e.g. texts and images) or multilingual settings (e.g. English and Chinese).

Also, we consider an interesting research direction in this area to be an aspect-based complaint classification task to identify the target of complaints in customer reviews, similar to aspect-based sentiment analysis (Pontiki et al., 2016).

- In Chapter 5, we identified the limitations of models in predicting bragging and their types. Future research can work on improving predictive performance by developing new methods and techniques. Also, the bragging data set introduced in Chapter 5 motivates future work to seek solutions to data imbalance (Li and Zhou, 2022).

- The task of identifying bragging and classifying their types presented in Chapter 5 can be extended to other bragging classification tasks, such as identifying different bragging strategies (e.g. explicit and implicit bragging, bragging as a complaint) (Dayter, 2014; Ren and Guo, 2020) and recognizing text-image incongruity in bragging (Matley, 2018).

- The work in Chapter 6 investigated bragging behaviors on Twitter in association with user socio-demographic traits. Similar methods can be used to explore the relationship between bragging and personalities (e.g. narcissism, openness) (Moon et al., 2016). Further studies can leverage larger and more diverse data sets to gain comprehensive insights into the complexities of bragging behavior across various online platforms and user demographics. Additionally, it can inspire large-scale empirical studies of other speech acts in computational linguistics from the perspective of users' personal traits.

Bibliography

Yasser Al-Shboul. 2021. Complaining Strategies by Jordanian Male and Female Students at BAU. *Dirasat, Human and Social Sciences*, 48(4).

Kristen Altenburger, Rajlakshmi De, Kaylyn Frazier, Nikolai Avteniev, and Jim Hamilton. 2017. Are There Gender Differences in Professional Self-Promotion? An Empirical Case Study of Linkedin Profiles Among Recent MBA Graduates. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Sarthak Anand, Debanjan Mahata, Kartik Aggarwal, Laiba Mehnaz, Simra Shahid, Haimin Zhang, Yaman Kumar, Rajiv Ratn Shah, and Karan Uppal. 2019. Suggestion Mining from Online Reviews using ULMFit. *arXiv preprint arXiv:1904.09076*.

Tatiana Anikina and Ivana Kruijff-Korbayova. 2019. Dialogue act classification in team communication for robot assisted disaster response. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 399–410, Stockholm, Sweden. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Norman Au, Dimitrios Buhalis, and Rob Law. 2009. Complaints on the Online Environment—The Case of Hong Kong Hotels. *Information and communication technologies in tourism 2009*, pages 73–85.

John L Austin. 1962. How to dothings with words. *Harvard University*.

JinYeong Bak, Suin Kim, and Alice Oh. 2012. Self-Disclosure and Relationship Strength in Twitter Conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 60–64.

JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing Twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1986–1996.

Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378.

David Bamman and Noah A Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *Proceedings of the 9th International Conference on Weblogs and Social Media*, ICWSM, pages 574–577.

Elizabeth Barrett and Vic Lally. 1999. Gender differences in an on-line learning environment. *Journal of computer assisted learning*, 15(1):48–60.

Enkhbold Bataa and Joshua Wu. 2019. An Investigation of Transfer Learning-Based Sentiment Analysis in Japanese. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4652–4657. Association for Computational Linguistics.

Roy F Baumeister. 1982. A self-presentational view of social phenomena. *Psychological bulletin*, 91(1):3.

Natalya N Bazarova, Jessie G Taft, Yoon Hyung Choi, and Dan Cosley. 2013. Managing Impressions and Relationships on Facebook: Self-Presentational and Relational Concerns Revealed Through the Analysis of Language Style. *Journal of Language and Social Psychology*, 32(2):121–141.

Tilman Beck, Ji-Ung Lee, Christina Viehmann, Marcus Maurer, Oliver Quiring, and Iryna Gurevych. 2021. Investigating label suggestions for opinion mining in German Covid-19 social media. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–13, Online. Association for Computational Linguistics.

Shreesh Kumara Bhat and Aron Culotta. 2017. Identifying Leading Indicators of Product Recalls from Online Reviews Using Positive Unlabeled Learning and Domain Adaptation. In *Eleventh International AAAI Conference on Web and Social Media*.

Rohan Bhatia, Apoorva Singh, and Sriparna Saha. 2022. Complaint and Severity Identification from Online Financial Content.

Diana Boxer. 1993a. *Complaining and commiserating: A speech act view of solidarity in spoken American English*. Lang.

Diana Boxer. 1993b. Social distance and speech behavior: The case of indirect complaints. *Journal of pragmatics*, 19(2):103–125.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge University Press.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Laura E Buffardi and W Keith Campbell. 2008. Narcissism and social networking web sites. *Personality and social psychology bulletin*, 34(10):1303–1314.

D. John Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1301–1309.

Lei Cao, Huijun Zhang, and Ling Feng. 2020. Building and Using Personal Knowledge Graph to Improve Suicidal Ideation Detection on Social Media. *IEEE Transactions on Multimedia*.

Angela L Carey, Melanie S Brucks, Albrecht CP Küfner, Nicholas S Holtzman, Mitja D Back, M Brent Donnellan, James W Pennebaker, Matthias R Mehl, et al. 2015. Narcissism and the use of personal pronouns revisited. *Journal of personality and social psychology*, 109(3):e1.

Rich Caruana. 1997. Multitask Learning. *Machine learning*, 28(1):41–75.

Azizeh Chalak. 2021. Pragmatics of Self-Praise and Self-Presentation by Iranian EFL Learners on Instagram. *TESL-EJ*, 25(1):n1.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. *arXiv preprint arXiv:2010.02559*.

Xi Chen, Gang Li, YunDi Hu, and Yujie Li. 2016. How Anonymity Influence Self-disclosure Tendency on Sina Weibo: An Empirical Study. *The anthropologist*, 26(3):217–226.

Zhenpeng Chen, Xuan Lu, Sheng Shen, Wei Ai, Xuanzhe Liu, and Qiaozhu Mei. 2017. Through a Gender Lens: An Empirical Study of Emoji Usage over Large-Scale Android Users. *arXiv preprint arXiv:1705.05546*.

Dhivya Chinnappa and Eduardo Blanco. 2018. Mining Possessions: Existence, Type and Temporal Anchors. 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 496–505, New Orleans, Louisiana. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint arXiv:1406.1078*.

Hui-Tzu Grace Chou and Nicholas Edge. 2012. "they are happier and having better lives than i am": The impact of using facebook on perceptions of others' lives. *Cyberpsychology, behavior, and social networking*, 15(2):117–121.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Mike Conway and Daniel O'Connor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology*, 9:77–82.

Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine learning*, 20(3):273–297.

Kristof Coussement and Dirk Van den Poel. 2008. Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision support systems*, 44(4):870–882.

Thomas Cover and Peter Hart. 1967. Nearest Neighbor Pattern Classification. *IEEE transactions on information theory*, 13(1):21–27.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.

Daria Dayter. 2014. Self-praise in microblogging. *Journal of Pragmatics*, 61:91–102.

Daria Dayter. 2018. Self-praise online and offline: The hallmark speech act of social media? *Internet Pragmatics*, 1(1):184–203.

Daria Dayter. 2021. Dealing with interactionally risky speech acts in simultaneous interpreting: The case of self-praise. *Journal of Pragmatics*, 174:28–42.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *Proceedings of the international conference recent advances in natural language processing ranlp 2013*, pages 198–206.

Daantje Derks, Arjan ER Bos, and Jasper Von Grumbkow. 2008. Emoticons and Online Message Interpretation. *Social Science Computer Review*, 26(3):379–388.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Steve Duck. 1998. *Human Relationships*. Sage.

Yuksel Ekinci, Joana Calderon, and Haytham Siala. 2016. Do personality traits predict 'complaining'consumers? *Journal of Business Environment*, 8(1):32.

Paul Ekman. 1992. An Argument for Basic Emotions. *Cognition & Emotion*, 6(3-4):169–200.

Ming Fang, Shi Zong, Jing Li, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2022. Analyzing the Intensity of Complaints on Social Media. *arXiv preprint arXiv:2204.09366*.

Elise Fehn Unsvåg and Björn Gambäck. 2018. The Effects of User Features on Twitter Hate Speech Detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 75–85, Brussels, Belgium. Association for Computational Linguistics.

Akash Gautam, Debanjan Mahata, Rakesh Gosangi, and Rajiv Ratn Shah. 2020. Semi-Supervised Iterative Approach for Domain-Specific Complaint Detection in Social Media. In *Proceedings of The 3rd Workshop on e-Commerce and NLP*, pages 46–53, Seattle, WA, USA. Association for Computational Linguistics.

David Gefen and Catherine M Ridings. 2005. If you spoke as she does, sir, instead of the way you do: A sociolinguistics perspective of gender differences in virtual communities. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, 36(2):78–92.

Ronald Geluykens and Bettina Kraft. 2007. Gender variation in native and interlanguage complaints. *Cross-cultural pragmatics and interlanguage English. Muenchen: LincomGmbH*.

Vahid Ghahraman and Mahboube Nakhle. 2013. A Contrastive Study on the Complaint Behavior among Canadian Native. *The Iranian EFL Journal*, 20(21.67):313.

David C Gilmore and Gerald R Ferris. 1989. The effects of applicant impression management tactics on interviewer judgments. *Journal of management*, 15(4):557–564.

Erving Goffman. 1967. Interaction ritual: Essays on face-to-face interaction.

Erving Goffman et al. 1978. *The Presentation of Self in Everyday Life*, volume 21. Harmondsworth London.

Yueguo Gu. 1990. Politeness phenomena in modern chinese. *Journal of pragmatics*, 14(2):237–257.

D Gunawan, RP Siregar, RF Rahmat, and A Amalia. 2018. Building automatic customer complaints filtering application based on Twitter in Bahasa Indonesia. In *Journal of Physics: Conference Series*, volume 978, page 012119. IOP Publishing.

Sharath Chandra Guntuku, Weisi Lin, Jordan Carpenter, Wee Keong Ng, Lyle H Ungar, and Daniel Preoţiuc-Pietro. 2017. Studying Personality through the Content of Posted and Liked Images on Twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 223–227.

Yaping Guo and Wei Ren. 2020. Managing image: The self-praise of celebrities on social media. *Discourse, Context & Media*, 38:100433.

Daniel Halpern, James E Katz, and Camila Carril. 2017. The online ideal persona vs. the jealousy effect: Two explanations of why selfies are associated with lower-quality romantic relationships. *Telematics and Informatics*, 34(1):114–123.

Jianmin He, Mengna Hu, Mingguang Shi, and Yezheng Liu. 2014. Research on the measure method of complaint theme influence on online social network. *Expert Systems with Applications*, 41(13):6039–6046.

Trine Heinemann and Véronique Traverso. 2009. Complaining in interaction.

Robert K Herbert. 1990. Sex-based differences in compliment behavior1. *Language in society*, 19(2):201–224.

John P Hewitt and Randall Stokes. 1975. Disclaimers. *American sociological review*, pages 1–11.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-Based Sentiment Analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland. Linköping University Electronic Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Robert Hogan. 1982. A socioanalytic theory of personality. In *Nebraska symposium on motivation*. University of Nebraska Press.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4405–4414, Brussels, Belgium. Association for Computational Linguistics.

Janet Holmes. 2013. *Women, Men and Politeness*. Routledge.

Thomas Holtgraves. 1990. The language of self-disclosure.

Vera Hoorens, Mario Pandelaere, Frans Oldersma, and Constantine Sedikides. 2012. The Hubris Hypothesis: You Can Self-Enhance, But You'd Better Not Show It. *Journal of personality*, 80(5):1237–1274.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 752–762.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics.

Nan Hu, Ting Zhang, Baojun Gao, and Indranil Bose. 2019. What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*, 72:417–426.

Kazuhiro Ito, Taichi Murayama, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. Identifying A Target Scope of Complaints on Social Media. In *The 11th International Symposium on Information and Communication Technology*, pages 111–118.

Kokil Jaidka, Salvatore Giorgi, H Andrew Schwartz, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2020. Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, 117(19):10165–10171.

Kokil Jaidka, Sharath Guntuku, and Lyle Ungar. 2018. Facebook versus Twitter: Differences in Self-Disclosure and Trait Prediction. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Jiahua Jin, Xiangbin Yan, You Yu, and Yijun Li. 2013. Service Failure Complaints Identification in Social Media: A Text Classification Approach.

Mali Jin and Nikolaos Aletras. 2020. Complaint Identification in Social Media with Transformer Networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1765–1771, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mali Jin and Nikolaos Aletras. 2021. Modeling the Severity of Complaints in Social Media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2264–2274, Online. Association for Computational Linguistics.

Mali Jin, Daniel Preotiuc-Pietro, A. Seza Doğruöz, and Nikolaos Aletras. 2022. Automatic Identification and Classification of Bragging in Social Media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3945–3959, Dublin, Ireland. Association for Computational Linguistics.

Edward E Jones. 1990. *Interpersonal perception.* WH Freeman/Times Books/Henry Holt & Co.

Edward E Jones, Thane S Pittman, et al. 1982. Toward a general theory of strategic self-presentation. *Psychological perspectives on the self*, 1(1):231–262.

Andi Kaharuddin. 2020. The Speech Act of Complaint: Socio-Cultural Competence Used by Native Speakers of English and Indonesian. *International Journal of Psychosocial Rehabilitation*, 24(06).

Leila Nasiri Kakolaki and Mohsen Shahrokhi. 2016. Gender Differences in Complaint Strategies among Iranian Upper Intermediate EFL Students.

Jihen Karoui, Farah Benamara Zitoune, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich Belguith. 2015. Towards a Contextual Pragmatic Model to Detect Irony in Tweets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 644–650, Beijing, China. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980.*

Taiwo Kolajo, Olawande Daramola, Ayodele Adebiyi, and Aaditeshwar Seth. 2020. A framework for pre-processing of social media feeds based on integrated local knowledge base. *Information Processing & Management*, 57(6):102348.

Robin M Kowalski. 1996. Complaints and complaining: Functions, antecedents, and consequences. *Psychological bulletin*, 119(2):179.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

M Lailiyah, S Sumpeno, and IK E Purnama. 2017. Sentiment Analysis of Public Complaints Using Lexical Resources Between Indonesian Sentiment Lexicon and Sentiwordnet. In *2017 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 307–312. IEEE.

Janice Laksana and Ayu Purwarianti. 2014. Indonesian Twitter text authority classification for government in Bandung. In *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 129–134. IEEE.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterising user impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 405–413, Gothenburg, Sweden. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*.

Mark R Leary and Robin M Kowalski. 1990. Impression management: A literature review and two-component model. *Psychological bulletin*, 107(1):34.

Marius Leckelt, David Richter, Carsten Schröder, Albrecht CP Küfner, Markus M Grabka, and Mitja D Back. 2019. The rich are different: Unravelling the perceived and self-reported personality profiles of high-net-worth individuals. *British Journal of Psychology*, 110(4):769–789.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Geoffrey Leech. 2016. *Principles of pragmatics*. Routledge.

Xiang Li and Yucheng Zhou. 2022. Disentangled and Robust Representation Learning for Bragging Classification in Social Media. *arXiv preprint arXiv:2210.15180*.

Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu. 2020. Generating Informative Conversational Response using Recurrent Knowledge-Interaction and Knowledge-Copy. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 41–52, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sarah MacIsaac, John Kelly, and Shirley Gray. 2018. 'she has like 4000 followers!': the celebrification of self within school social networks. *Journal of Youth Studies*, 21(6):816–835.

Carmen Maíz-Arévalo. 2021. "Blowing our own trumpet": Self-praise in Peninsular Spanish face-to-face communication. *Journal of Pragmatics*, 183:107–120.

Daniel N Maltz and Ruth A Borker. 2018. A cultural Approach to Male-Female Miscommunication. In *The matrix of language*, pages 81–98. Routledge.

Antonis Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020. Analyzing Political Parody in Social Media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4373–4384, Online. Association for Computational Linguistics.

David Matley. 2018. "this is Not a# humblebrag, this is just a# brag": The pragmatics of self-praise, hashtags and politeness in Instagram posts. *Discourse, context & media*, 22:30–38.

David Matley. 2020. Isn't working on the weekend the worst?# humblebrag": The impact of irony and hashtag use on the perception of self-praise in Instagram posts.

Christian Maurer and Sabrina Schaich. 2011. Online customer reviews used as complaint management tool. In *ENTER*, pages 499–511.

Francis T McAndrew and Hye Sun Jeong. 2012. Who does what on Facebook? Age, sex, and relationship status as predictors of Facebook use. *Computers in human behavior*, 28(6):2359–2365.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling Framing in Immigration Discourse on Social Media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. *arXiv preprint arXiv:1708.02182*.

Minas Michikyan, Jessica Dennis, and Kaveri Subrahmanyam. 2015. Can You Guess Who I Am? Real, Ideal, and False Self-Presentation on Facebook among Emerging Adults. *Emerging Adulthood*, 3(1):55–64.

Lynn Carol Miller, Linda Lee Cooke, Jennifer Tsang, and Faith Morgan. 1992. Should I Brag? Nature and Impact of Positive and Boastful Disclosures for Women and Men. *Human Communication Research*, 18(3):364–399.

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

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, *SEM, pages 1–17.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a Word–Emotion Association Lexicon. *Computational intelligence*, 29(3):436–465.

Jang Ho Moon, Eunji Lee, Jung-Ah Lee, Tae Rang Choi, and Yongjun Sung. 2016. The role of narcissism in self-promotion on Instagram. *Personality and individual Differences*, 101:22–25.

Kyunghye Moon. 2001. Speech act study: Differences between native and nonnative speaker complaint strategies. *The American University*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Sapna Negi, Tobias Daudert, and Paul Buitelaar. 2019. SemEval-2019 Task 9: Suggestion Mining from Online Reviews and Forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 877–887, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Beth Murphy—Joyce Neu. 1996. My grade's too low: The speech act set of complaining. *Speech acts across cultures: Challenges to communication in a second language*, 11:191.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200*.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. 2016. Computational Sociolinguistics: A Survey. *Computational linguistics*, 42(3):537–593.

Nhung Thi-Hong Nguyen, Phuong Phan-Dieu Ha, Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Vietnamese Complaint Detection on E-Commerce Websites. *arXiv preprint arXiv:2104.11969*.

Wantira Noisiri. 2002. Speech act of complaint: Pragmatic study of complaint behaviour between males and females in thai. *University of Sussex*, 1:18.

Elite Olshtain and Liora Weinbach. 1987. Complaints: A Study of Speech Act Behavior among Native and Non-native Speakers of Hebrew. In *The pragmatic perspective*, page 195. John Benjamins.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A Dataset of Intended Sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.

Nikolaos Panagiotou, Ioannis Katakis, and Dimitrios Gunopulos. 2016. Detecting Events in Online Social Networks: Definitions, Trends and Challenges. *Solving Large Scale Learning Tasks. Challenges and Algorithms*, pages 42–84.

Widya Paramita and Felix Septianto. 2021. The benefits and pitfalls of humblebragging in social media advertising: the moderating role of the celebrity versus influencer. *International Journal of Advertising*, pages 1–24.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Barbara Plank and Dirk Hovy. 2015. Personality Traits on Twitter—or—How to Get 1,500 Personality Tests in a Week. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 92–98.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Damar Adi Prabowo and Guntur Budi Herwanto. 2019. Duplicate question detection in question answer website using convolutional neural network. In *2019 5th International Conference on Science and Technology (ICST)*, volume 1, pages 1–6. IEEE.

Marília Prada, David L Rodrigues, Margarida V Garrido, Diniz Lopes, Bernardo Cavalheiro, and Rui Gaspar. 2018. Motives, frequency and attitudes toward emoji and emoticon use. *Telematics and Informatics*, 35(7):1925–1934.

Justin P Preddie and Monica Biernat. 2021. More than the Sum of Its Parts: Intersections of Sexual Orientation and Race as They Influence Perceptions of Group Similarity and Stereotype Content. *Sex Roles*, 84(9-10):554–573.

Daniel Preotiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. 2016. Studying the Dark Triad of Personality through Twitter Behavior. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 761–770.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically Identifying Complaints in Social Media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying User Income through Language, Behaviour and Affect in Social Media. *PloS one*, 10(9):e0138717.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint Modelling of Emotion and Abusive Language Detection. *arXiv preprint arXiv:2005.14028*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC, pages 37–44.

Leonard Reinecke and Sabine Trepte. 2014. Authenticity and well-being on social network sites: A two-wave longitudinal study on the effects of online authenticity and the positivity bias in sns communication. *Computers in Human Behavior*, 30:95–102.

Wei Ren and Yaping Guo. 2020. Self-praise on Chinese social networking sites. *Journal of Pragmatics*, 169:179–189.

Wei Ren and Yaping Guo. 2021. What is "Versailles Literature"?: Humblebrags on Chinese social networking sites. *Journal of Pragmatics*, 184:185–195.

Devan Rosen, Michael A Stefanone, and Derek Lackaff. 2010. Online and offline social networks: Investigating culturally-specific behavior and satisfaction. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, *SEM, pages 502–518.

Sofia Rüdiger and Daria Dayter. 2020. Manbragging online: Self-praise on pick-up artists' forums. *Journal of Pragmatics*, 161:16–27.

Jian Rui and Michael A Stefanone. 2013. Strategic self-presentation online: A cross-cultural study. *Computers in human behavior*, 29(1):110–118.

H Sacks. 1992. In jefferson g.(ed.), 5 lectures on conversation.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7685–7697, Online. Association for Computational Linguistics.

Roger C Schank. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive psychology*, 3(4):552–631.

Barry R Schlenker. 1980. *Impression management*, volume 222.

Barry R Schlenker and Mark R Leary. 1982. Audiences' reactions to self-enhancing, self-denigrating, and accurate self-presentations. *Journal of experimental social psychology*, 18(1):89–104.

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. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PloS one*, 8(9):e73791.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential Language Analysis Toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60.

Irene Scopelliti, George Loewenstein, and Joachim Vosgerau. 2015. You call it "Self-Exuberance"; I call it "Bragging" miscalibrated predictions of emotional responses to self-promotion. *Psychological science*, 26(6):903–914.

Kate Scott. 2015. The pragmatics of hashtags: Inference and conversational style on Twitter. *Journal of Pragmatics*, 81:8–20.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

John R Searle. 1976. A Classification of Illocutionary Acts. *Language in society*, 5(1):1–23.

Constantine Sedikides. 1993. Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of personality and social psychology*, 65(2):317.

Constantine Sedikides, Aiden P. Gregg, and Claire M. Hart. 2007. The importance of being modest. *The Self: Frontiers in Social Psychology, edited by Constantine Sedikides and Stephen J. Spencer*, 163(84).

Ovul Sezer, Francesca Gino, and Michael I Norton. 2018. Humblebragging: A distinct—and ineffective—self-presentation strategy. *Journal of Personality and Social Psychology*, 114(1):52.

Taha Shangipour ataei, Soroush Javdan, and Behrouz Minaei-Bidgoli. 2020. Applying Transformers and Aspect-based Sentiment Analysis approaches on Sarcasm Detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 67–71, Online. Association for Computational Linguistics.

Pavica Sheldon. 2013. Examining Gender Differences in Self-disclosure on Facebook Versus Face-to-Face. *The Journal of Social Media in Society*, 2(1).

Robert F Simmons. 1965. Answering English questions by computer: a survey. *Communications of the ACM*, 8(1):53–70.

Apoorva Singh, Soumyodeep Dey, Anamitra Singha, and Sriparna Saha. 2022a. Sentiment and Emotion-aware Multi-modal Complaint Identification.

Apoorva Singh, Arousha Nazir, and Sriparna Saha. 2022b. Adversarial Multi-task Model for Emotion, Sentiment, and Sarcasm Aided Complaint Detection. In *European Conference on Information Retrieval*, pages 428–442. Springer.

Apoorva Singh and Sriparna Saha. 2021. Are you really complaining? a multi-task framework for complaint identification, emotion, and sentiment classification. In *International Conference on Document Analysis and Recognition*, pages 715–731. Springer.

Apoorva Singh, Sriparna Saha, Md Hasanuzzaman, and Kuntal Dey. 2022c. Multitask learning for complaint identification and sentiment analysis. *Cognitive Computation*, 14(1):212–227.

Apoorva Singh, Sriparna Saha, Mohammed Hasanuzzaman, and Anubhav Jangra. 2021a. Identifying complaints based on semi-supervised mincuts. *Expert Systems with Applications*, 186:115668.

Apoorva Singh, Tanmay Sen, Sriparna Saha, and Mohammed Hasanuzzaman. 2021b. Federated Multi-task Learning for Complaint Identification from Social Media Data. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 201–210.

Raghvendra Pratap Singh, Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. 2020. Identifying Complaints from Product Reviews: A Case Study on Hindi.

Susan A Speer. 2012. The Interactional Organization of Self-praise: Epistemics, Preference Organization, and Implications for Identity Research. *Social Psychology Quarterly*, 75(1):52–79.

Himani Srivastava, Vaibhav Varshney, Surabhi Kumari, and Saurabh Srivastava. 2020. A Novel Hierarchical BERT Architecture for Sarcasm Detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 93–97, Online. Association for Computational Linguistics.

R Sujay, Jagadeesh Pujari, Vandana Shreenivas Bhat, and Anita Dixit. 2018. Timeline Analysis of Twitter User. *Procedia computer science*, 132:157–166.

Didi Sukyadi et al. 2011. Complaining in EFL Learners: Differences of Realizations between Men and Women (A Case Study of Indonesian EFL Learners at the English Department of

The Indonesian University of Education). *PAROLE: Journal of Linguistics and Education*, 2(1 April):1–25.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Sharyl Tanck. 2002. Speech act sets of refusal and complaint: A comparison of native and non-native english speakers' production. *Studies in Second Language Acquisition*, 13(1):65–81.

Philip E Tetlock. 2002. Social functionalist frameworks for judgment and choice: intuitive politicians, theologians, and prosecutors. *Psychological review*, 109(3):451.

Zhengxi Tian, Wenge Rong, Libin Shi, Jingshuang Liu, and Zhang Xiong. 2018. Attention Aware Bidirectional Gated Recurrent Unit Based Framework for Sentiment Analysis. In *International Conference on Knowledge Science, Engineering and Management*, pages 67–78. Springer.

Dianne M Tice, Jennifer L Butler, Mark B Muraven, and Arlene M Stillwell. 1995. When modesty prevails: Differential favorability of self-presentation to friends and strangers. *Journal of personality and social psychology*, 69(6):1120.

Suhatati Tjandra, Amelia Alexandra Putri Warsito, and Judi Prajetno Sugiono. 2015. Determining citizen complaints to the appropriate government departments using KNN algorithm. In *2015 13th International Conference on ICT and Knowledge Engineering (ICT & Knowledge Engineering 2015)*, pages 1–4. IEEE.

Els Tobback. 2019a. On downgrading and upgrading strategies used in the act of self-praise in French and US LinkedIN-summaries. A contrastive pragmatic analysis. *Social Media Corpora for the Humanities (CMC-Corpora2019)*, page 7.

Els Tobback. 2019b. Telling the world how skilful you are: Self-praise strategies on Linkedin. *Discourse & Communication*, 13(6):647–668.

Anna Trosborg. 2011. *Interlanguage pragmatics: Requests, complaints, and apologies*, volume 7. Walter de Gruyter.

Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata. 2018. Nowcasting the Stance of social media users in a Sudden Vote: The Case of the Greek Referendum.

In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 367–376.

Prasanna Umar, Chandan Akiti, Anna Squicciarini, and Sarah Rajtmajer. 2021. Self-disclosure on Twitter during the COVID-19 Pandemic: A Network Perspective. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part IV 21*, pages 271–286. Springer.

Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2019. Detection and Analysis of Self-Disclosure in Online News Commentaries. In *The World Wide Web Conference*, pages 3272–3278.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2022. A Multi-task Model for Sentiment Aided Stance Detection of Climate Change Tweets. *arXiv preprint arXiv:2211.03533*.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying Medical Self-Disclosure in Online Communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4398–4408, Online. Association for Computational Linguistics.

Silvie Válková. 2013. Speech acts or speech act sets: Apologies and compliments. *Linguistica Pragensia*, 2(23):44–57.

Carolien Van Damme, Eliane Deschrijver, Eline Van Geert, and Vera Hoorens. 2017. When Praising Yourself Insults Others: Self-Superiority Claims Provoke Aggression. *Personality and Social Psychology Bulletin*, 43(7):1008–1019.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Camilla Vásquez. 2011. Complaints Online: The Case of TripAdvisor. *Journal of Pragmatics*, 43(6):1707–1717.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Svitlana Volkova and Yoram Bachrach. 2015. On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychology, Behavior, and Social Networking*, 18(12):726–736.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring Perceived Demographics from User Emotional Tone and User-Environment Emotional Contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1567–1578.

Jun Wang, Kelly Cui, and Bei Yu. 2021. Self Promotion in US Congressional Tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4893–4899, Online. Association for Computational Linguistics.

Kai Wang and Tat-Seng Chua. 2010. Exploiting Salient Patterns for Question Detection and Question Retrieval in Community-based Question Answering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1155–1163.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Anchalee Wannaruk. 2008. Pragmatic Transfer in Thai EFL Refusals. *RELC journal*, 39(3):318–337.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Liyun Wen, Xiaojie Wang, Zhenjiang Dong, and Hong Chen. 2017. Jointly Modeling Intent Identification and Slot Filling with Contextual and Hierarchical Information. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 3–15. Springer.

Joanna Wolfe and Elizabeth Powell. 2006. Gender and Expressions of Dissatisfaction: A Study of Complaining in Mixed-Gendered Student Work Groups. *Women and Language*, 29:2.

Ruey-Jiuan Regina Wu. 2011. A conversation analysis of self-praising in everyday Mandarin interaction. *Journal of Pragmatics*, 43(13):3152–3176.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510.

Wei Yang, Luchen Tan, Chunwei Lu, Anqi Cui, Han Li, Xi Chen, Kun Xiong, Muzi Wang, Ming Li, Jian Pei, and Jimmy Lin. 2019a. Detecting Customer Complaint Escalation with Recurrent Neural Networks and Manually-Engineered Features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 56–63, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, pages 5754–5764.

Joanna C Yau and Stephanie M Reich. 2019. "It's Just a Lot of Work": Adolescents' Self-Presentation Norms and Practices on Facebook and Instagram. *Journal of research on adolescence*, 29(1):196–209.

Zhen Zuo, Bing Shuai, Gang Wang, Xiao Liu, Xingxing Wang, Bing Wang, and Yushi Chen. 2016. Learning Contextual Dependence with Convolutional Hierarchical Recurrent Neural Networks. *IEEE Transactions on Image Processing*, 25(7):2983–2996.

Appendix A

Guidelines of Complaint Severity Annotation

Text classification models for complaint identification are computer systems that get as input a piece of text and return a prediction which is typically a choice from a predefined set of categories (labels). For example, given the text *Our room was a rip-off. We were paying almost $300 a night and expected luxury*, it can be classified according to its context into one of two categories: **complaint** or **non-complaint**.

Complaining is a speech act extensively used by humans to communicate a negative inconsistency between reality and expectations. Based on severity level, complaints are classified into four categories: (a) **No Explicit Reproach**; (b) **Disapproval**; (c) **Accusation**; (d) **Blame** (see the definitions below). Given the complaint *I love <user>! Shame you're introducing a man tax of 7% in 2018 :(*, it can be classified according to its severity level into one of these four categories.

Definitions

- **Example 1** in Table A.1 implies dissatisfaction instead of directly complaining or mentioning the cause and is therefore classified into "No Explicit Reproach".
- **Example 2** in Table A.1 expresses obvious negative sentiment (*horrible*) and is therefore classified into "Disapproval".
- **Example 3** in Table A.1states that someone bumped into the car (accuses someone

Category	**Definition**	**Example**
No Explicit Reproach	Complainer does not explicitly mention the cause of the dissatisfaction in the complaint and does not directly state something offensive	**Example 1:** *My car was in perfect order when I last drove it. There was nothing wrong with my car yesterday.*
Disapproval	Complainer expresses dislike, disapproval, and annoyance in connection with a certain state of affairs that he or she considers bad for them	**Example 2:** *There's a horrible dent in my car. Oh dear, I've just bought it.*
Accusation	Complainer establishes the complainee as the agent of the problem and directly or indirectly accuses the complainee for causing the problem but there is no direct blame	**Example 3:** *You borrowed my car last night, didn't you?* (Indirect) *Did you happen to bump into my car?* (Direct)
Blame	Complainer assumes that the complainee is guilty of the offence and states modified blame of complainee's action or directly blames the complainee on his or her action	**Example 4:** *Honestly, couldn't you have been more careful? You should take more care with other people's car.*

Table A.1: Definitions and examples of complaint severity levels.

of something) and is therefore classified into "Accusation".

- **Example 4** in Table A.1 points out that someone is responsible for crashing the car (blame someone for something) and is therefore classified into "Blame".

Accusations vs. Blame Formally, we use accuse with an act—we assert that somebody did something reprehensible: *Mary accuses John of failing to lock the door.* But we use blame with the outcome-we assert somebody's responsibility for the undesirable result: *Mary blames John for the robbery.* OR *Mary blames the robbery on John.*

Step

You have to decide which severity level the complaint belongs to according to the above definition.

You should choose one of the following options:

- **No Explicit Reproach** (No explicit mention of the cause, not offensive)
- **Disapproval** (express dissatisfaction, annoyance, dislike, disapproval)
- **Accusation** (direct or indirect accusation of the complainee)
- **Blame** (assumes that the complainee is responsible, contains directly/indirectly blame)

Note: If a complaint expresses disapproval and accusation or disapproval and blame then it should be labelled as either accusation or blame.

Example

The following text is a complaint:

I love <user>! Shame you're introducing a man tax of 7% in 2018 :(

Question : Choose the severity level of the complaint:

- **No Explicit Reproach**
- **Disapproval**
- **Accusation**
- **Blame**

The answer is Blame

Example of reasoning: The complainer emphasizes that the complainee (i.e., <*user*>) is responsible for introducing a man tax.

Appendix B

Guidelines and Interface for Bragging Annotation

Thank you for your participation in our study. During our experiment, we will ask you to read and evaluate a tweet which may include a bragging or a praisal statement.

Instructions You need to identify whether or not a tweet includes a bragging statement.

Bragging Bragging is a speech act which explicitly or implicitly attributes credit to the speaker for some 'good' (possession, accomplishment, skill, etc.) which is positively valued by the speaker and the potential audience. As such, bragging includes announcements of accomplishments, explicit positive evaluations of some aspect of self and other types defined below. A bragging statement should clearly express what the author is bragging about (i.e. the target of bragging).

If the tweet is about bragging, decide on the category where the tweet belongs to from the following categories:

Achievement The act of bragging is about a concrete outcome obtained as a result of the tweet author's actions. These results may include achievements, awards, products, and/or positive change in a situation or status (individually or as part of a group).

Examples:

- *Finally got that offer! Whoop!!*
- *Our team won the championship*

Action The act of bragging is about a past, current or upcoming action of the user that does not have a concrete outcome

Examples:

- *Hanging at Buffalo Wild Wings with @user for the #ILLvsASU game. #BragginGRights*
- *Guess what! I met Matt Damon today!*

Feeling The act of bragging is about a feeling that is expressed by the user for a particular situation.

Example:

- *Im so excited that I am back on my consistent schedule. I am so excited for a routine so I can achieve my goals!!*

Trait The act of bragging is about a personal trait, skill or ability of the user .

Examples:

- *To be honest, I have a better memory than my siblings*
- *I look great after losing weight*

Possession The act of bragging is about a tangible object belonging to the user.

Example:

- *Look at our Christmas tree! I kinda just wanna keep it up all year!*

Affiliation The act of bragging is about being part of a group (e.g. family, team, org etc.) and/or a certain location including living in a city, neighborhood or country, enrolled into a university, supporting a team, working in a company etc.

Example:

- *My daughter got first place in the final exam, so proud of her!*

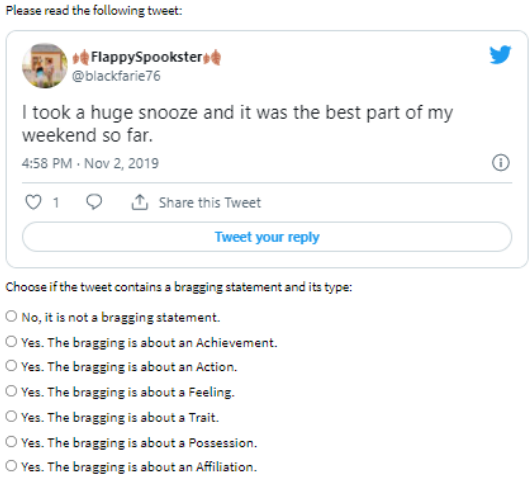
Not bragging If the tweet is not about bragging, then select "No. This is not a bragging statement."

Examples:

- *One of the best books I've ever read*
- *hahahahahaha*
- *You gotta admit, that's some mighty awesome aim!*
- *Vote in the poll below for your book of choice!*
- *I think this is great*
- *dear everyone announcing they are at "Friendsgiving", we get it, you have friends*
- *In case you didn't know, Adam Silver is in charge*
- *I feel terrible*
- *I don't know why you are celebrating*
- *This is exactly what is going on!*
- *I love you*

Select "No. This is not a bragging statement", also in cases when:

- there is not enough information to determine that the tweet is about bragging
- the bragging statements belong to someone other than the author of the tweet



Please read the following tweet:

FlappySpookster
@blackfarie76

I took a huge snooze and it was the best part of my weekend so far.

4:58 PM · Nov 2, 2019

Choose if the tweet contains a bragging statement and its type:

○ No, it is not a bragging statement.
○ Yes. The bragging is about an Achievement.
○ Yes. The bragging is about an Action.
○ Yes. The bragging is about a Feeling.
○ Yes. The bragging is about a Trait.
○ Yes. The bragging is about a Possession.
○ Yes. The bragging is about an Affiliation.

Figure B.1: Screenshot of annotation interface on our platform.

- the relationship between author and people/things mentioned in the tweet are unknown:
  - *This kid is smart*
  - *That was an amazing stream*
  - *Kudos to mike Dunleavy! It's hard to get a franchise record ANYTHING in Chicago*
- the post is about the act of bragging:
  - *We want to hear you brag!*
  - *Trump isn't Bragging anymore as his tradewar hits the stockmarket hard*
  - *Dudes are getting too cocky these days. Them lil labels and that dar don't impress everyone. brag differently*

Not available Finally, if the tweet is not available or displayed, or is in a language other than English, please select the ”Not available” option.

Other considerations Please verify the content of hashtags as these may give clues towards the category of the tweet. The judgment should be made only based on the given content of the tweet - please do not search the tweet on Twitter or online in order to identify additional context.