NII IDR 2021

複数コーパスへの尤度分布による意見文の性質表現

P17

山西良典（関西大学），竹元亨舟（関西大学），西原陽子（立命館大学），吉田光男（筑波大学）

謝辞：本研究は，NII公募型共同研究戦略研究公募型の支援のもと実施した．また，国立情報学研究所のIDRデータセット提供サービスにより株式会社Insight Tech から提供を受けた「不満調査データセット」，弁護士ドットコム株式会社から提供を受けた「弁護士ドットコムデータセット」，および，「千葉大学 3人会話コーパス」を利用した

本研究の貢献

アノテーションデータを使わずに意味のあるテキスト分類の実現

既存コーパスの**新しい応用先**

ラベルの分布表現化による意見文の詳細な性質表現

背景と問題

機械学習を用いたテキスト分類はアノテーションデータが必要

アノテーションデータの有無が研究対象になるかどうか？！

既存コーパスの有効利用で解決できないか！？

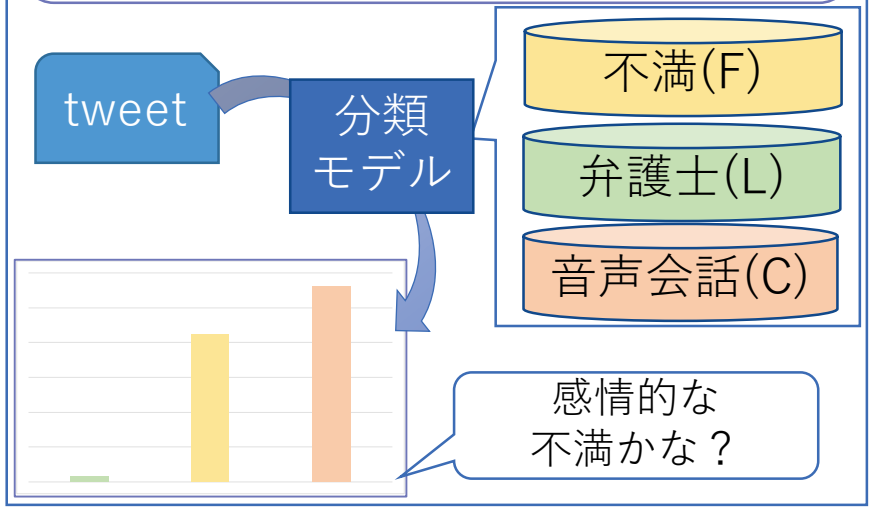
提案手法

1．既存コーパスへの分類器を学習

・分析対象のデータにアノテートして学習データは作らない
・既存コーパスの種類ごとにデータを分類するように学習

2．分類結果の尤度を解釈

・複数コーパスそれぞれへの尤度分布を考察
・分布形状をもとに性質を表現



実験と考察

・緊急事態宣言が発令された2021/08/11の10:00~23:00に政治家公式アカウントへリプライされたtweet集合を分析
・モデルは音声チャット(C)，法律相談(L)，不満調査の不満意見(F)のコーパス分類をfastTextで学習

尤度分布とtweetの傾向

Cのみ高：短文，内容が少ないコメント
Lのみ高：丁寧な表現が含まれる
Fのみ高：丁寧ではない表現が含まれる

L&F高：丁寧な表現で不満を表現．相手が聞く気になりそう
F&C高：内容のない言葉で不満を表現．相手に届かない言葉の可能性
L&C高：ほぼなし．これらは共存しない!?