https://doi.org/10.17323/jle.2024.22443

Обнаружение поддерживающих высказываний (Hope Speech) с использованием дискурса социальных сетей (Posi-Vox-2024): подход на основе обучения с переносом

Мухаммад Ахмад [1], Сардар Усман [2], Хумайра Фарид [3], Икра Амир [4], Мухаммад Музамил[5], Хмаза Амир [5], Григорий Сидоров [1], Ильдар Батыршин [1]

[1] Национальный политехнический институт (CIC-IPN), Мехико, Мексика

[2] Институт искусств и культуры, Лахор, Пакистан

[3] Независимый исследователь, Калифорния, США

[4] Университет штата Пенсильвания в Абингтоне, Пенсильвания, США

[5] Исламский университет Бахавалпура, Пакистан

АННОТАЦИЯ

Введение: Понятие надежды определяется как оптимистичное ожидание или предвкушение положительных результатов. В эпоху активного использования социальных сетей исследования в основном сосредоточены на монолингвальных подходах, при этом языки урду и арабский остаются недостаточно изученными.

Цель: Данное исследование посвящено совместному многоязычному обнаружению поддерживающих высказываний на урду, английском и арабском языках с использованием парадигмы обучения с переносом. Мы создали новый многоязычный набор данных под названием Posi-Vox-2024 и применили совместную многоязычную технику для разработки универсального классификатора, подходящего для многоязычных данных. Мы протестировали дообученную модель BERT, которая продемонстрировала высокую эффективность в распознавании семантической и контекстной информации.

Методология: Структура включает (1) предварительную обработку, (2) представление данных с использованием BERT, (3) дообучение и (4) классификацию поддерживающих высказываний в бинарные («надежда» и «не надежда») и многоклассовые (реалистичные, нереалистичные и обобщенные надежды) категории.

Результаты: Предложенная нами модель (BERT) показала наивысшую производительность на нашем наборе данных, достигнув точности 0,78 в бинарной классификации и 0,66 в многоклассовой классификации. Это улучшило результаты на 0,04 и 0,08 соответственно по сравнению с базовыми показателями (логистическая регрессия: 0,75 для бинарной и 0,61 для многоклассовой классификации).

Заключение: Наши результаты могут быть использованы для улучшения автоматизированных систем обнаружения и продвижения поддерживающего контента на английском, арабском и урду на платформах социальных сетей, способствуя формированию позитивного онлайн-дискурса. Это исследование устанавливает новые стандарты для обнаружения многоязычных поддерживающих высказываний, расширяя существующие знания и открывая возможности для будущих исследований на недостаточно изученных языках.

КЛЮЧЕВЫЕ СЛОВА

высказывание надежды, BERT, машинное обучение, Twitter[1], социальные сети, обучение с переносом, обработка естественного языка.

Для цитирования: Ахмад, М., Усман, С., Фарид, Х., Амир, И., Музамил, М., Амир, Х., Сидоров, Г., & Батыршин, И. (2024). Обнаружение поддерживающих высказываний (Hope Speech) с использованием дискурса социальных сетей (Posi-Vox-2024): Подход на основе обучения с переносом. *Journal of Language and Education, 10*(4), 31-44. https://doi.org/10.17323/jle.2024.22443

Контакты:
Ильдар Батыршин
batyr1@cic.ipn.mx

Получена: 1 ноября 2024

Принята: 16 декабря 2024

Опубликована: 30 декабря 2024

ВВЕДЕНИЕ

Надежда определяется как позитивное эмоциональное состояние, включающее ожидания или предвкушение благоприятных событий в будущем. Многие онлайн-платформы социальных сетей предоставили пространство для миллионов

[1] Ресурс заблокирован на территории Российской Федерации по решению Роскомнадзора.

пользователей, где они могут выражать свои мысли и делиться своими взглядами. Эта возможность не только привела к появлению негативного контента, но и способствовала обмену позитивными идеями (Alawadh et al., 2023), а также продвижению позитивного контента. В последнее время выявление высказываний, выражающих надежду, в социальных сетях привлекает значительное внимание, однако лишь небольшое число исследований затрагивает эту проблему применительно к языкам с большим и малым количеством ресурсов (Arif et al., 2024; Balouchzahi et al., 2023; Chakravarthi, 2022). Выявление высказываний надежды — это относительно новый подход, направленный на выявление и усиление позитивного онлайн-контента с целью продвижения социальной гармонии и создания более позитивной атмосферы в сообществах. Среди немногочисленных исследований по выявлению высказываний надежды основное внимание уделялось одноязычным контекстам, где разрабатывались индивидуальные классификационные модели для каждого языка, включая английский (Balouchzahi et al., 2023), испанский (Kumar et al., 2022), английский, тамильский и малаялам (RamakrishnaIyer et al., 2023), а также бенгальский (Nath et al., 2023), в то время как арабский язык и урду не рассматривались ни в одноязычном, ни в многоязычном контексте.

Для многих людей социальные сети стали жизненно важной платформой для поиска поддержки (Gowen et al., 2012; Yates et al., 2017; Wang & Jurgens, 2018). Социальная интеграция имеет решающее значение для их общего благополучия, особенно для тех, кто подвержен риску социальной изоляции. Обнаруживая и выделяя ободряющие сообщения в социальных сетях, выявление высказываний надежды может способствовать созданию более справедливого и всеохватывающего цифрового пространства. Кроме того, разработанная в этом исследовании методология может применяться в психолингвистике и обработке естественного языка для выявления позитивных настроений, психологической устойчивости и конструктивного дискурса в различных контекстах.

Платформы социальных медиа содержат множество враждебных или злонамеренных сообщений (Louati, Ali, et al., 2024; Irfan, Asim, et al., 2024; Anjum, and Rahul Katarya., 2024), во многом из-за отсутствия регуляции. Анализ контента в Твиттере и на других платформах показал свою эффективность в сдерживании распространения негатива с помощью таких методов, как выявление высказываний вражды (Schmidt & Wiegand, 2017; Subramanian, Malliga, et al., 2023; Nagar, Barbhuiya, & Dey, 2023), оскорблений (Anand et al., 2023; Kogilavani et al., 2023; Mnassri et al., 2024) и агрессии (Zampieri et al., 2019; Austin et al., 2020; Yenala et al., 2018). Тем не менее, как показывают последние исследования, существующие технологии выявления агрессивных высказываний (Lee et al., 2018) часто не учитывают потенциальные искажения, присутствующие в наборах данных, на которых они обучаются. Наличие систематических расовых предубеждений в этих данных может сделать алгоритмы выявления агрессивных высказываний изначально предвзятыми, что ведет к дискриминационным последствиям, особенно в отношении меньшинств или маргинализированных групп. Такие искажения в технологиях обработки языка могут способствовать закреплению дискриминации (Davidson et al., 2019). Поэтому следует уделять приоритетное внимание продвижению позитивных взаимодействий, а не ограничиваться только реакцией на отдельные негативные публикации. В этом контексте выявление высказываний надежды предлагает новый подход, который не только противодействует негативу, но и способствует созданию более позитивной и инклюзивной онлайн-среды в широком спектре лингвистических и культурных контекстов. Для достижения этой цели мы создали комплексный совместный многоязычный корпус с высказываниями надежды на урду, арабском и английском языках с использованием бинарной и многоклассовой классификации.

Процесс начинается со сбора данных, связанных с твитами, содержащими высказывания надежды на английском, урду и арабском языках из Твиттера. После сбора данных каждый образец проходит предварительную подготовку для повышения устойчивости к работе моделей машинного обучения. Затем данные проходят процесс совместной многоязычной обработки, где объединяются наборы данных на урду, английском и арабском языках. На этапе аннотирования данные маркируются согласно определенным инструкциям. Следующим шагом является дообучение предложенных моделей с последующим применением к набору данных для задач классификации. Наконец, различные модели машинного обучения, глубокого обучения и трансформерные модели оцениваются по метрикам Accuracy, F1-score, Recall и Precision, и результаты анализируются для задач бинарной и многоклассовой классификации. Такая методология обеспечивает всесторонний подход к выявлению высказываний надежды на разных языках.

Получены следующие результаты:

(1) Насколько нам известно, методы совместного многоязычного выявления высказываний надежды для урду, английского и арабского языков ранее не разрабатывались. Мы создали комплексный совместный многоязычный корпус и предоставили подробные рекомендации по аннотированию набора данных.
(2) Мы рассматривали проблему выявления высказываний надежды как задачу двухуровневой классификации текста в совместном многоязычном наборе данных (английский, арабский и урду) и предложили метод многоклассовой классификации для урду и арабского языка.
(3) Комплексная серия экспериментов показала, что предложенная методология достигла лучших результатов по сравнению с базовым уровнем.
(4) Предложенная модель продемонстрировала значение Accuracy 0,78 в двоичной классификации и 0,66 в многоклассовой классификации для нашего набора данных. Это представляет собой улучшение Accuracy на

0,04 в бинарной классификации и 0,08 в многоклассовой классификации по сравнению с базовыми показателями качества.

ЛИТЕРАТУРНЫЙ ОБЗОР

Существующие наборы данных для выявления высказываний надежды

Процесс создания корпусов для выявления высказываний надежды стал важным направлением в этой области, однако такие корпуса, как правило, ограничены как по языковому охвату, так и по объему выборки. Например, Balouchzahi et al. (2023) недавно представили набор данных для выявления высказываний надежды на английском языке и провели сравнительный анализ с использованием методов машинного обучения, глубокого обучения, а также обучения на основе трансформеров. Тем не менее, этот набор данных был ограничен одним языком, и в исследовании не рассматривалась многоязычная классификация. Аналогично Chakravarthi (2022) представил модель CNN для выявления высказываний надежды на английском и дравидийских языках, но не рассматривал аспекты многоязычной классификации. Эти исследования подчеркивают необходимость создания более разнообразных наборов данных, включающих несколько языков, это способствует улучшению обобщения. В частности, Chakravarthi (2022) создал объединенный многоязычный набор данных для английского, тамильского и малаяламского языков, используя комментарии с YouTube с целью распознавания и поощрения позитивности в комментариях, однако автор не использовал задачу многоязычной классификации для выявления высказываний надежды.

Многоязычное выявление высказываний надежды

Несколько исследований посвящены многоязычному выявлению высказываний надежды, в них авторы применяли современные передовые модели машинного обучения для учета лингвистических и культурных различий между языками. К примеру, Ghanghor et al. (2021) использовали предобученные трансформерные модели, такие как m-BERT-cased и XLM-RoBERTa, для выявления высказываний надежды на английском, тамильском и малаяламском языках. Их результаты показали, что m-BERT-cased работает лучше всех остальных моделей, достигая максимального значения F1-score: 0,93 для английского, 0,83 для малаяламского и 0,60 для тамильского. Несмотря на значимый вклад в многоязычное выявление высказываний надежды, в этой работе не исследуется задача многоклассовой классификации на разных языках. В свою очередь, Chinnappa (2021) также занимался выявлением высказываний надежды на тамильском, английском и малаяламском языках, отмечая сложности, вызванные данными со смешанным кодом, что дополнительно усложняет задачу классификации. В дальнейшем Malik et al. (2023) расширили область исследования, применив совместный многоязычный и основанный на переводе подход, сосредоточив внимание на английском и русском языках, что подчеркивает потенциал методов перевода в многоязычном выявлении высказываний надежды. Они дообучили предобученную модель Russian-RoBERTa и достигли впечатляющих результатов с Accuracy 94 % и F1-score 80,24 %. Этот подход продемонстрировал потенциал использования перевода для улучшения про-

Таблица 1

Сравнение предыдущих исследований, связанных с выявлением высказываний надежды, с предлагаемым исследованием

Библиографическая ссылка	Язык	Совместный многоязычный	Обучение с учителем	Мультиклассовая классификация
Balouchzahi et al. (2023)	English	No	LR, SVM, CNN, LSTM, BiLSTM, Transformer	Yes
Malik et al. (2024)	English, Russian	Yes	SVM, RF, CNN, RoBET base with classifier	No
Kumar et al. (2022)	English, Spanish, Tamil, Malayalam	No	SVM, LR, RF	No
Roy et al. (2022)	English	No		No
Chakravarthi et al. (2021)	English, Tamil, Malayalam, Kannada	No	SVM, DT, LR, KNN, RoBERTa Classifier	No
Ghanghor et al. (2021)	English	No		No
Proposed	English, Urdu, Arabic	Yes	DT, CatBoost, XGB, LR, BiLSTM, CNN, BGRU, BERT, DistilBERT	Yes

изводительности модели, однако задача многоклассовой классификации осталась нерешенной и представляет собой важное направление для дальнейших исследований.

Вклад данного исследования

Наше исследование представляет новый подход, сосредоточенный на совместном многоязычном выявлении высказываний надежды на урду английском и арабском языках с использованием двухуровневой текстовой классификации. В отличие от предыдущих исследований, которые фокусировались на отдельных языках, наша методология предлагает комплексный совместный многоязычный набор данных, включающий как бинарную, так и многоклассовую задачи классификации. Данное исследование вносит значимый вклад в выявление высказываний надежды на трех языках, открывая новые возможности для улучшения анализа настроений и инструментов мониторинга социальных медиа. В Таблице 1 представлен обзор предыдущих исследований, связанных с выявлением высказываний надежды, в котором выделены отличия между ними и предлагаемым исследованием.

МЕТОДОЛОГИЯ

Корпус

Сбор и интеграция набора данных

Наш набор данных включает приблизительно 80 000 твитов из различных дисциплин на урду, английском и арабском языках, полученных из социальной сети «Твиттер» — крупнейшей платформе микроблогов, позволяющей публиковать сообщения, известные как «твиты», и взаимодействовать с пользователями. Сбор данных осуществлялся с помощью API Twitter[2] (Tweepy). В этом исследовании мы собрали корпус из 80 000 недавних твитов, отобранных на основе ключевых слов, которые были получены из Твиттера. Для фоомирования этого корпуса мы применили систематический подход, используя ключевые слова, связанныес надеждой, например, в урду: انشاء اللہ (Ин Ша Аллах), ریخ انشاء اللہ (Хайр Ин Ша Аллах), ہے شہ اوخ (желание), لک (завтра), لبقتسم (будущее), یبایماک (успех), راظتنا (ожидание), روپر ھب ے س دیما (надеющийся) и т. д., тогда как в английском языке мы использовали (стремление, вера, скоро, мечта, ожидание, позитивный настрой, желаю, с нетерпением жду и радостный) и т. д., а в арабском языке мы использовали لؤافتلا (оптимизм), تشجيع (поощрение) , باهذ , ةقفاوم (одобрение), يتمنى (желание) и т. д. с различными вариациями. Эти ключевые слова использовались для охвата разнообразного спектра выражений и настроений, присутствующих на платформе. Сбор данных осуществлялся с сентября 2023 г. по март 2024 г., что обеспечило прочную основу для проведения углубленного анализа и исследования динамики коммуникации, связанной с выражением надежды в цифровом пространстве. После сбора образцов из Твиттера мы объединили наши данные на английском, арабском и урду в один CSV-файл, получивший название Posi-Vox-2024. Термин «Posi Vox», образованный от «Posi» (позитивный) и «Vox» (голос), фокусируется на высказываниях, выражающих надежду, и направлен на выявление позитивного дискурса в многоязычных сообществах. На Рисунке 1 представлены предлагаемая нами методология и дизайн исследования, описывающие процесс анализа высказываний надежды в смешанных текстах, часто встречающихся в обсуждениях в социальных сетях в многоязычных сообществах. Термин «многоязычное выявление высказываний надежды» относится к этому унифицированному подходу, который обрабатывает и интерпретирует смешанные тексты для улучшения анализа тональности в различных онлайн-сообществах. Разработанная модель учитывает лингвистические нюансы без перевода, что делает её чрезвычайно актуальной для многоязычных социальных сетей. Она предлагает масштабируемое решение для обнаружения риторики надежды в смешанном контенте, обеспечивая большую гибкость и надежность по сравнению с традиционными одноязычными моделями, тем самым улучшая анализ тональности и способствуя позитивному дискурсу в различных онлайн-сообществах.

Предварительная обработка данных

API Tweepy[3] был разработан для предоставления функций фильтрации твитов по различным критериям, таким как дата, местоположение, язык и идентификатор твита. В частности, мы использовали атрибуты даты и языка для сбора твитов на урду, английском и арабском языках. В связи с большим количеством шума в текстовом контенте социальных сетей мы провели ряд процедур предварительной обработки:

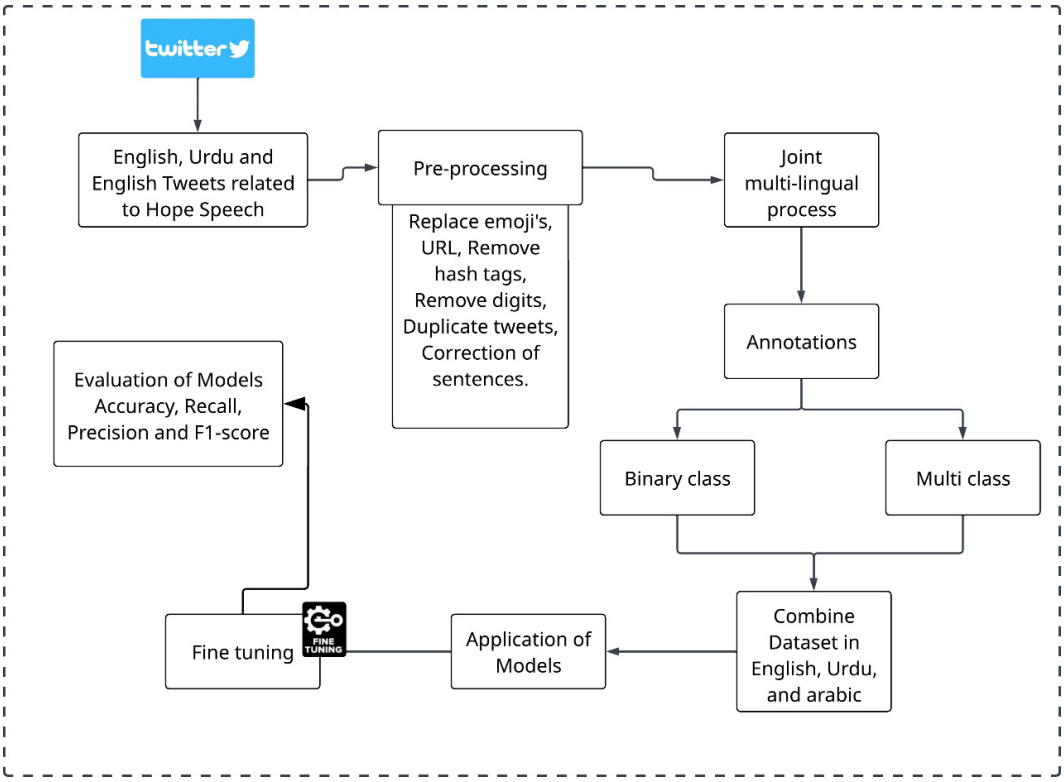
(1.) Исключение URL-адресов, упоминаний пользователей в форме @use и HTML-тегов.
(2.) Удаление знаков препинания из текста.
(3.) Удаление дубликатов и твитов длиной менее 20 символов.
(4.) Преобразование заглавных букв в строчные.
(5.) Замена эмодзи соответствующим текстом; как мы знаем, эмодзи играют важную роль в распознавании твитов.
(6.) Удаление цифр из твитов.
(7.) Расшифровка всех коротких текстов, таких как thnx для Thanks, plz для Please и т. д.

[2] Ресурс заблокирован на территории Российской Федерации по решению Роскомнадзора.

[3] https://www.tweepy.org/ Последнее посещение: 11.10.2024.

Рисунок 1

Предлагаемая методология и дизайн



После обработки 80 000 твитов осталось только 18 362 оригинальных твита, содержащих текст на урду, английском и арабском языках. Они были использованы для создания совместного многоязычного набора данных.

Процесс аннотации

Правила аннотирования

Основываясь на определении надежды, данном психологами, мы разделили твиты на два класса. К основному классу относились твиты, выражающие надежду, а ко второму - твиты, лишенные какого-либо чувства надежды. Эта методология классификации позволяет нам анализировать и интерпретировать наличие или отсутствие надежды в содержании твитов, что, в свою очередь, помогает глубже понять настроения и эмоциональные проявления пользователей в социальных сетях. На следующем этапе анализа мы классифицировали твиты по различным типам надежды, изучая специфические особенности и характеристики, присутствующие в их содержании. Мы разработали конкретные рекомендации для первичной и вторичной категоризации твитов, которые подробно описаны вместе с примерами в Таблицах 2 и 3.

NHS: Твит не передает никакого чувства надежды, стремления, желания или предвкушения будущего.

(1.) Обобщенная надежда: форма надежды, характеризующаяся общим чувством оптимизма и позитивного настроения, которое не связано с каким-либо конкретным событием или результатом.

(2.) Нереалистичная надежда: категория, для которой свойственно присутствие ожидания некого события, несмотря на низкую или практически отсутствующую вероятность этого. Иногда люди могут питать надежду на иррациональные события или исходы, такие ожидания могут быть вызваны сильными эмоциями — гневом, печалью или депрессией.

(3.) Реалистичная надежда: этот тип надежды предполагает ожидание чего-то разумного, значимого и находящегося в пределах возможного. При этом существует высокая вероятность того, что ожидаемые события действительно произойдут.

Отбор аннотаторов

Мы намеренно избегали отбора аннотаторов для набора данных Posi-Vox -2024 на основе расовой принадлежности, тем самым продемонстрировав нашу непоколебимую приверженность продвижению культуры равенства и разнообразия, сохраняя при этом целостность набора данных. Мы предприняли целенаправленные усилия по учету национальности аннотаторов, избегая при этом учета расовой принадлежности. Такой подход позволил нам беспристрастно отслеживать географическое разнообразие наших

Таблица 2

Высказывания надежды бинарного класса

№.	Твиты	Категория
1	جب کسی کام کا ارادہ کر لیں تو مکمل اللہ پہ بھروسہ کریں اور ہر شک و شبہ Если вы хотите, чтобы вы знали, как это сделать, вы можете сделать это в любое время. بہترین کارساز ہے (Когда вы намереваетесь что-то сделать, полностью положитесь на Аллаха и удалите все сомнения из своего сердца, потому что Он — лучший исполнитель, которому вы доверили свои дела.)	Надеяться
2	Всем все равно, ты никому не нужен, а эти нестандартные чернокожие мужики тебя раскручивают	Не надеяться

Таблица 3

Высказывания надежды для всех классов

№	Твиты	Категория
1	Встречайте каждый день с оптимизмом, верьте в светлое завтра и доверяйте предстоящему пути	Обобщенная надежда
2	Верьте, что ваша жизнь наладится, и всё будет хорошо. Верьте, что ваше здоровье улучшится, а любовь придет к вам	Реалистичная надежда
3	Я мечтаю летать без крыльев, парить над облаками, бросая вызов гравитации	Нереалистичная надежда

Таблица 4

Аннотаторы на основе географического региона

Язык	Страна	Мужской	Женский	Бакалавриат	Аспирантура
Английский	Великобритания	2	0	2	0
урду	Пакистан	2	1	2	1
арабский	ОАЭ	2	0	2	0

аннотаторов, как показано в Таблице 4, без учета каких-либо предубеждений, связанных с расовыми характеристиками.

Процедура аннотирования

Выбранным аннотаторам были предоставлены подробные рекомендации по аннотированию и примеры аннотаций в разделе «Настройка аннотаций». Все аннотаторы, перечисленные в Таблице 4, обладают сильными навыками аннотирования, они имеют как степени бакалавра, так и магистра, а также значительный опыт в области обработки естественного языка, машинного обучения и глубокого обучения. Для контроля процесса аннотирования были созданы индивидуальные Google-формы для каждого аннотатора, запланированы еженедельные встречи для оценки прогресса аннотирования и выявления любых трудностей, возникших в ходе процесса. На Рисунке 2 показаны этапы создания корпуса для выявления высказываний надежды в твитах социальных сетей. Изначально набор данных проходит бинарную классификацию, чтобы отличить твиты, которые содержат признаки надежды, от твитов, не демонстрирующих их. Затем внутри положительного класса проводится более детальная классификация, в рамках которой выделяют конкретные эмоциональные категории, такие как обобщенная надежда, реалистичная надежда и нереалистичная надежда.

Статистика набора данных

На Рисунке 3 представлено облако слов, содержащее ключевые слова, извлеченные из твитов в многоязычном наборе данных, посвященном теме «высказывания надежды». На Рисунке 4 показано распределение меток как для бинарной, так и для многоклассовой классификации. Мы собрали равное количество данных, относящихся к категориям «надежда» и «не надежда», чтобы продемонстрировать баланс данных, также нам было необходимо дополнительно классифицировать категорию «надежда» по нескольким классам, таким как «обобщенная надежда», «реалистичная надежда» и «нереалистичная надежда», основываясь на эмоциях.

Ключевые характеристики набора данных о высказываниях надежды включают общее количество твитов (n = 18362), общий размер словарного запаса (n = 105777), общее количество слов (n = 499486), общее количество символов (n = 2 583 769), среднее количество слов (n = 27,32) и среднее количество символов (n = 141,56), как указано в Таблице 5.

Дополнение данных

Для повышения производительности и надежности предлагаемой модели мы применили метод обратного перевода с дополнениями к данным. В качестве инструмента обратного перевода мы использовали API Google Translate, который

Рисунок 2

Процедура аннотации при выявлении высказываний надежды

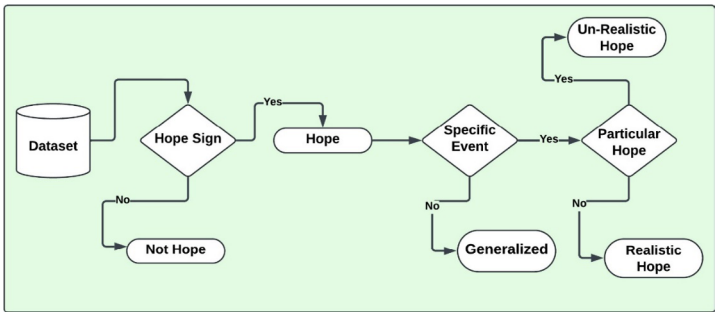


Рисунок 3

«Облако» слов надежды

Рисунок 4

Распределение меток бинарных и мультиклассов во всем наборе данных

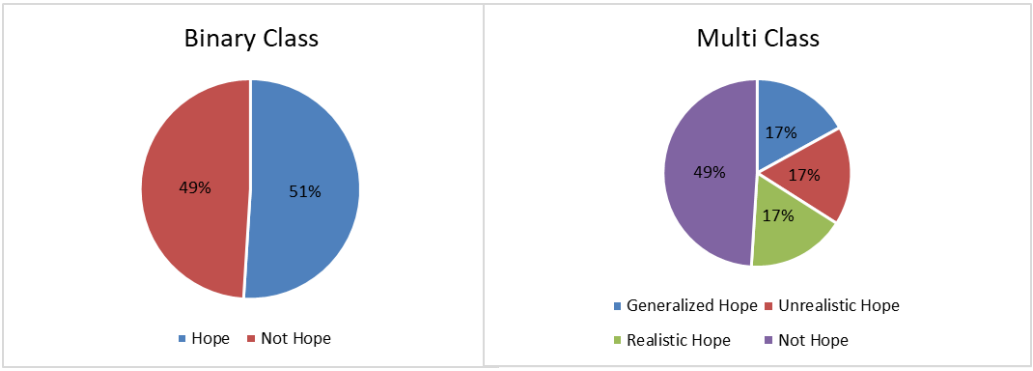


Таблица 5

Статистика набора данных

Классы	Твиты	Слова	Среднее число слов	Среднее число букв	Число букв	Размер словаря
Обобщенные надежды	3082	82046	26,61	139,39	429603	19617
Реалистичные надежды	3156	90355	28,63	150,12	473791	21286
Нереалистичные надежды	3074	83248	27,08	137,90	423933	20693
“Не надежда”	9050	243837	26,94	138,83	1256442	44181
всего	18362	499486	27,32	141,56	2583769	105777

отличается широким охватом языков и высоким качеством перевода. Для автоматизации процесса перевода и эффективной обработки больших объемов текста были разработаны специальные скрипты. После обратного перевода мы провели ручную проверку качества выборки дополненных данных, чтобы убедиться в сохранении смысла исходного текста и отсутствии значительной потери информации при переводе.

Этическая предосторожность

Данные, собранные в Твиттере, являются конфиденциальными, и мы уделяем особое внимание мерам по обеспечению их безопасности в процессе аннотирования. Личности участников процесса оставались скрытыми, а наши аннотаторы, выявляя имена политиков или знаменитостей, придерживались протокола невмешательства и не предпринимали попыток установить с ними контакт.

Методы выявления высказываний надежды

Для демонстрации того, как предложенный нами корпус Posi-Vox-2024 может быть использован для разработки, оценки и сравнения методов выявления высказываний надежды, мы применили и сравнили четыре современные модели машинного обучения: (i) Decision Tree (DT), (ii) CatBoost (CB), (iii) Extreme Gradient Boosting (XGB) и (iv) Logistic Regression (LR); три модели глубокого обучения: (i) Bidirectional Long Short-Term Memory (BiLSTM), (ii) Convolutional Neural Network (CNN) и (iii) Bidirectional Gated Recurrent Unit (BGRU); а также две модели трансферного обучения: (i) pre-trained BERT и (ii) distilBERT.

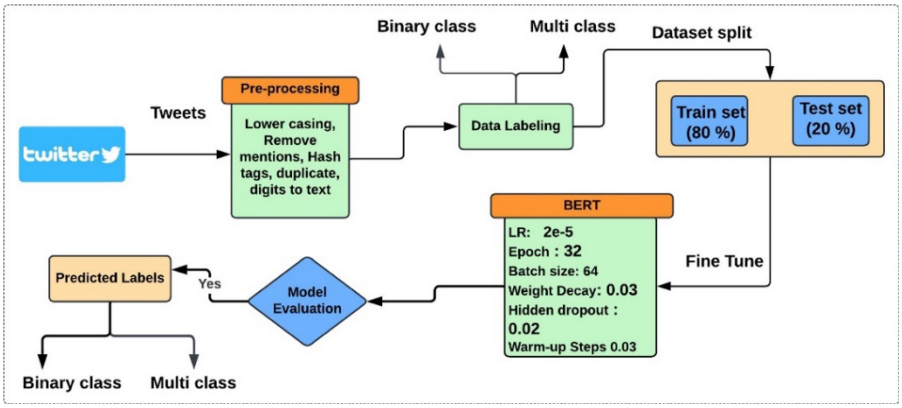
Наша лучшая модель основана на архитектуре BERT, использующей слои трансформера для захвата контекстных взаимосвязей в многоязычном тексте. После предварительной обработки с применением соответствующих токенизаторов для урду, английского и арабского языков, мы дообучили предобученную модель BERT на нашем размеченном наборе данных с использованием функции потерь cross-entropy и оптимизатора Adam со скоростью обучения 2e-5. Набор данных был разделен: 80 % для обучения и 20 % для тестирования. Для обеспечения воспроизводимости конфигурации в Таблице 9 представлены: размер батча, количество эпох, значения метрик оценки, а также оптимальные значения гиперпараметров. На Рисунке 5 показана диаграмма, иллюстрирующая архитектуру BERT и поток данных, она необходима для пояснения процесса обработки многоязычного ввода и предсказания высказываний надежды.

РЕЗУЛЬТАТЫ

В этом разделе представлены результаты применения различных моделей машинного обучения, глубокого обучения и трансформеров для задачи многоязычного выявления высказываний надежды. Эти модели оценивались на задачах бинарной и многоклассовой классификации с использованием предложенного нами корпуса Posi-Vox-2024, включающего тексты на английском и арабском языках, а также урду. В Таблицах 6, 7, 8, 10 и 11 приведены результаты по метрикам Precision, Recall, F1-score и Accuracy, полученные с помощью современных алгоритмов машинного обучения, таких как Decision Tree (DT), Categorical Boosting (CatBoost), Extreme Gradient Boosting (XGB) и Logistic Regression (LR). Для глубокого обучения мы использовали модели Convolutional Neural Network (CNN), Bidirectional Gated Recurrent Unit (BGRU) и Bidirectional Long Short-Term Memory (BiLSTM). На нашем корпусе Posi-Vox-2024 в категории трансформеров применялись модели Bidirectional Encoder Representations from Transformers (BERT) и Distilled BERT (DistilBERT). Целью наших экспериментов был поиск наиболее подходящей модели для выявления высказываний надежды на разных языках, при этом для каждой модели систематически проводилась настройка гиперпараметров и анализировалась эффективность на основе таких

Рисунок 5

Конвейер обучения модели на основе BERT для классификации многоязычных текстов



метрик, как Acuracy, Precision, Recall и F1-score. В последующих подразделах приводятся подробные результаты для каждой категории моделей.

Машинное обучение

Таблица 6 демонстрирует результаты, достигнутые различными моделями машинного обучения с использованием TF-IDF-векторизации для задачи выявления высказываний надежды в рамках бинарной и многоклассовой классификации. Для бинарной классификации моделей DT, CatBoost, XGB и LR показатели F1-score варьируют от 0,70 до 0,73, при этом модель LR достигает наивысших значений Precision, Recall, F1-score и Accuracy — 0,75. В задаче многоклассовой классификации высказываний надежды модель LR также показала лучшие результаты, достигнув F1-score 0,61. Модели CatBoost и XGB продемонстрировали конкурентоспособные показатели с Accuracy на уровне 0,58. Таким образом, LR превосходит другие модели как в бинарной, так и в многоклассовой задачах, достигая наивысших значений Precision, Recall, F1-score и Accuracy.

Глубокое обучение

Таблица 7 содержит показатели эффективности трех моделей глубокого обучения (CNN, BGRU и BiLSTM) для бинарной и многоклассовой классификации. В рамках задачи бинарной классификации все три модели показывают схожие результаты: CNN и BiLSTM достигают Precision, Recall и F1-score на уровне 0,75, тогда как модель BGRU имеет немного меньшие значения - 0,74. Accuracy для всех трех моделей также стабильна и составляет 0,75 для CNN и BiLSTM и 0,74 для BGRU. В многоклассовой классификации наблюдается снижение показателей по всем метрикам. Модели CNN и BGRU показывают Precision, Recall, F1-Score и Accuracy около 0,56, тогда как BiLSTM демонстрирует несколько лучшие результаты — примерно 0,62 по всем метрикам. Это свидетельствует о том, что модели хорошо справляются с бинарной классификацией, однако они испытывают большие трудности при многоклассовой классификации.

Результаты трансформеров

В Таблице 8 приведены значения метрик качества Precision, Recall, F1-score и Accuracy для решения задач бинарной и многоклассовой классификации. Для бинарной классификации модель BERT демонстрирует более высокие значения Precision, Recall, F1-score и Accuracy — все по 0,78, по сравнению с DistilBERT, у которого показатели несколько ниже: F1-score 0,75 и Accuracy 0,76. При переходе к многоклассовой классификации эффективность обеих моделей снижается: BERT достигает F1-score 0,65 и Acuracy 0,66, тогда как DistilBERT имеет немного более низкие показатели F1-score и Accuracy — 0,64. В целом, модель BERT превосходит DistilBERT в обеих категориях задач.

Таблица 9 содержит оптимальные параметры дообучения предобученной модели BERT для решения задач бинарной и многоклассовой классификации. Лучшие гиперпараметры были подобраны с помощью grid search, учитывая сле-

Таблица 6

Результаты моделей машинного обучения

Классификация	Модель	Precision	Recall	F1-score	Accuracy
Бинарная классификация	DT	0,72	0,72	0,72	0,72
	Catboost	0,73	0,73	0,73	0,73
	XGB	0,72	0,71	0,7	0,71
	LR	0,75	0,75	0,75	0,75
Мультиклассовая классификация	DT	0,59	0,59	0,59	0,59
	Catboost	0,57	0,58	0,53	0,58
	XGB	0,57	0,58	0,54	0,58
	LR	0,61	0,61	0,61	0,61

Таблица 7

Результаты моделей глубокого обучения

Классификация	Модель	Precision	Recall	F1-score	Accuracy
Бинарная классификация	CNN	0,75	0,75	0,74	0,75
	BGRU	0,74	0,74	0,74	0,74
	BiLSTM	0,75	0,75	0,75	0,75
Мультиклассовая классификация	CNN	0,56	0,56	0,56	0,56
	BGRU	0,55	0,55	0,55	0,55
	BiLSTM	0,62	0,62	0,62	0,62

Таблица 8

Результаты трансформеров

Классификация	Модель	Precision	Recall	F1-score	Accuracy
Бинарная классификация	Bert	0,78	0,78	0,78	0,78
	DistilBert	0,76	0,76	0,75	0,76
Мультиклассовая классификация	Bert	0,66	0,66	0,65	0,66
	DistilBert	0,64	0,64	0,64	0,64

Таблица 9

Оптимальные значения гиперпараметров модели BERT

Гиперпараметры	Grid search
Learning rate	1e-5,1e-2, 2e-5, 3e-5, 3e-4
Эпохи	3, 9, 32
Размер батча	32, 64, 128
Weight Decay	0,01–0,1
Hidden dropout	0,02, 0,1
Warm-up Steps	0,03–0,1

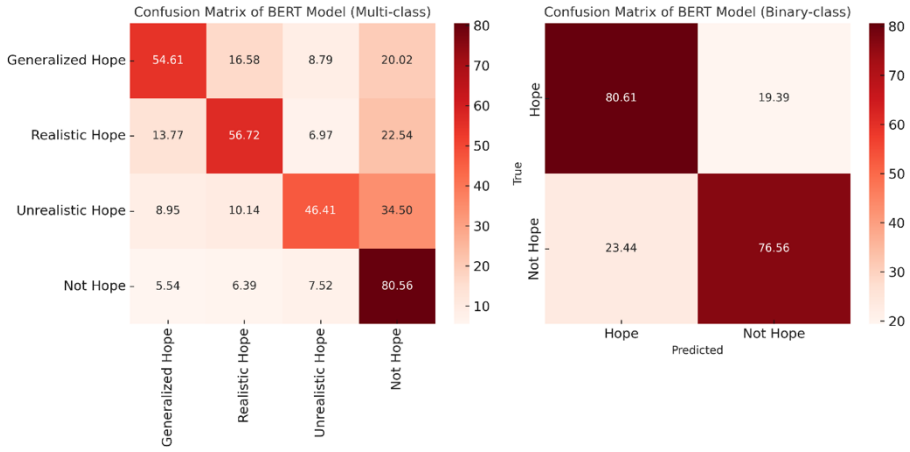
Таблица 10

Оценка по классам для предлагаемой методологии

Классификация	Категории	Precision	Recall	F1-Score	Support	Accuracy
Бинарная классификация	Надежда	0,78	0,79	0,78	3666	0,78
	Не надежда	0,79	0,78	0,78	3631	
Мультиклассовая классификация	Обобщенные надежды	0,50	0,63	0,55	1194	0,66
	Реалистичные надежды	0,57	0,53	0,55	1220	
	Нереалистичные надежды	0,63	0,41	0,50	1252	
	Не надежда	0,75	0,80	0,77	3631	

Рисунок 6

Матрица ошибок предлагаемой методологии



дующие диапазоны: learning rate — 1e-5, 1e-2, 2e-5, 3e-5, 3e-4; количество эпох — 9, 32, 64; размеры батча — 64, 128, 512; значения weight decay — от 0,01 до 0,1; коэффициенты скрытого dropout — 0,02 и 0,1; а также количество warm-up шагов — от 0,03 до 0,1. Эти настройки обеспечивают сбалансированную эффективность обучения и устойчивую производительность модели для различных задач классификации.

Анализ ошибок

В Таблице 10 представлены показатели по каждому классу, а на Рисунке 6 показаны матрицы ошибок (confusion matrix) в процентах для решения задач бинарной и многоклассовой классификации, достигнутых предложенной нами моделью. Отметим, что наша модель продемонстрировала лучшую точность (Precision) в классе Not Hope (не надежда). При классификации класса Unrealistic Hope (нереалистичная надежда) наша модель показывает лучшие результатОбобы по сравнению с другими метками класса Hope (надежда), при этом точность при определении категорий Generalized (обобщенная) и Realistic Hope (реалистичная надежда) ниже.

В Таблице 11 представлены результаты лучших моделей каждого подхода к обучению для бинарной и многоклассовой классификации. Предобученная модель BERT продемонстрировала заметное улучшение результатов по сравнению с традиционными моделями машинного обучения: для задачи бинарной классификации точность составила 0,78 против 0,75 у модели Logistic Regression (LR), что соответствует улучшению примерно на 4 %. В рамках задачи многоклассовой классификации BERT достиг точности 0,66, превосходя LR с показателем 0,61, что свидетельствует о повышении производительности примерно на 8,20 %. Это говорит о более глубокоТранм понимании контекста и способности учитывать тонкости языка, особенно в многоязычной среде. Таким образом, языковая модель BERT обеспечивает значительные преимущества в выявлении высказываний надежды на разных языках.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В исследовании рассматривается набор признаков, эффективно выявляющих высказывания надежды в твитах из Твиттера. Полученные результаты имеют важное значение для пользователей Интернета и общества в целом, способствуя укреплению мира и позитива. Надежда часто ассоциируется с поддержкой, ободрением, уверенностью, советами, вдохновением, которые необходимы людям в периоды болезни, стресса, одиночества или депрессии (Snyder et al., 2002). Обзор литературы показывает, что большинство исследований фокусировались преимущественно на бинарной классификации для выявления высказываний надежды в различных социальных сетях, при этом работа по созданию многоязычной системы была ограничена. Для устранения этого пробела мы разработали многоязычный инструмент, объединяющий совместную многоязычную методологию для решения задачи выявления высказываний надежды на урду, английском и арабском языках. Наш инструмент был обучен и протестирован на многоязычном наборе данных с целью выявления практических инсайтов и обеспечения применимости в реальном времени. Результаты показывают, что предложенный нами метод является эффективным и мощным инструментом для выявления высказываний надежды в твитах Твиттера. Мы использовали возможности трансферного обучения, дообучив предобученную модель BERT, это позволило нам достичь новыхого высот уровня точности: 78 % для бинарной классификации и 66 % для многоклассовой. Кроме того, наша методология превзошла четыре базовые модели: LR, XGB, CB и DT. Таким образом, на основании полученных результатов, предлагаемая нами система может быть применена для решения других задач многоязычной текстовой классификации в смежных областях.

В исследовании имеются несколько ограничений. Во-первых, сбор и аннотирование данных о высказываниях надежды на урду, английском и арабском языках сопряжены с рядом трудностей. Основной проблемой является поиск носителей языка, свободно владеющих данными языками и обладающих знаниями в области NLP и машинного обучения для обеспечения точной и надежной разметки. Во-вторых, при аннотировании встречались многочисленные твиты, выражающие надежду, но имеющие негативный подтекст. Например, на урду встречаются такие твиты, как

Таблица 11

Наиболее эффективные модели для каждого подхода к обучению

Классификация	Модель	Подход к обучению	Accuracy
Бинарная классификация	LR	Машинное обучение	0,75
	BiLSTM	Глубокое обучение	0,75
	BERT	Трансформер	**0,78**
Мультиклассовая классификация	LR	Машинное обучение	0,61
	BiLSTM	Глубокое обучение	0,62
	BERT	Трансформер	**0,66**

«میری امید ہے کہ میرے دشمنوں کو تباہی کا سامنا کرنا پڑے، ان کی تباہی میری خوشی ہوگی# USER-» («Моя надежда, что мои враги пострадают от разрушения, их разрушение будет моей радостью. #USER»), которые создают сложности, поскольку хотя твит передает надежду, основной его эмоциональный окрас негативен, что усложняет процесс разметки. В-третьих, урду и арабский языки считаются малоресурсными с точки зрения машинного и глубокого обучения, что затрудняет их понимание и обработку, и, как следствие, усложняет разработку устойчивых моделей для выявления высказываний надежды. В-четвертых, несмотря на высокую эффективность многоязычного выявления высказываний надежды, предлагаемая работа имеет ограничения. Корпус Posi-Vox-2024 ограничен по размеру и разнообразию, что влияет на обобщаемость результатов. Нюансы, связанные с языковой спецификой и смешанным кодом в данных, не были полностью учтены, что может негативно повлиять на точностьи классификации. Сложность модели и требования к ресурсам ограничивают её доступность, а производительность может снижаться со временем из-за динамичного характера дискурса в социальных медиа.

ЗАКЛЮЧЕНИЕ

Социальные сети стали мощной платформой для общественного диалога, оказывая влияние на мнения и эмоциональный фон сообществ. До настоящего времени большинство исследований было сосредоточено на противодействии негативным проявлениям в английском языке, в частности на выявлении высказываний ненависти. Данное исследование акцентирует основное внимание на важности многоязычного выявления высказываний надежды (Multilingual Hope Speech Detection, MHSD) в дискурсе социальных медиа, особо выделяя языки урду и арабский, которые ранее оставались вне внимания исследователей. Для достижения этой цели мы применили двухуровневую текстовую классификацию и создали комплексный набор данных под названием Posi-Vox-2024, включающий три языка — английский, урду и арабский. Такой подход позволил нам справиться с вызовами, связанными с многоязычностью, а также улучшить коммуникацию между представителями разных культур. Создавая многоязычный корпус и используя современные модели трансферного обучения с дообучением, мы успешно решили задачи выявления высказываний надежды на английском, арабском и урду. Результаты показывают, что предложенная нами система на основе предобученной модели BERT значительно превзошла четыре базовые модели (DT, XGB, CatBoost и LR), достигнув точности 0,78 для бинарной классификации и 0,66 для многоклассовой. Эти результаты подчеркивают важность продвижения позитивного дискурса в онлайн-среде и демонстрируют потенциал высказываний надежды как средства для формирования более здоровых и конструктивных взаимодействий в сообществах. Дальнейшие исследования могут быть направлены на расширение корпуса данных и включение дополнительных языков для повышения обобщаемости и устойчивости предлагаемой модели.

БЛАГОДАРНОСТИ

Работа выполнена при частичной поддержке правительства Мексики в рамках гранта A1-S-47854 CONAHCYT, а также грантов 20241816, 20241819, 20240936 и 20240951 Национального политехнического института Мексики. Авторы благодарят CONAHCYT за предоставленные вычислительные ресурсы через платформу глубокого обучения для языковых технологий Лаборатории суперкомпьютерных технологий INAOE, Мексика, также выражают благодарность компании Microsoft за поддержку в рамках программы Microsoft Latin America PhD Award.

ДОСТУПНОСТЬ ДАННЫХ

Данные будут предоставлены по запросу.

ВКЛАД АВТОРОВ

Мухаммад Ахмад: концептуализация; курирование данных; методология; ресурсы; программное обеспечение; визуализация; написание –первоначальный черновик; написание – рецензирование и редактирование.

Усман Сардар: курирование данных; формальный анализ; методология.

Хумайра Фарид: исследование; визуализация; написание – рецензирование и редактирование.

Икра Амир: курирование данных; формальный анализ; методология; программное обеспечение; написание – первоначальный черновик; написание – рецензирование и редактирование.

Мухаммад Музамил: курирование данных; методология; программное обеспечение.

Амир Хмаза: курирование данных; методология.

Григорий Сидоров: концептуализация; ресурсы; руководство; валидация.

Ильдар Батыршин: концептуализация; администрирование проекта; ресурсы; руководство; валидация; написание – первоначальный черновик; написание – рецензирование и редактирование.

ЛИТЕРАТУРА

Alawadh, H. M., Alabrah, A., Meraj, T., & Rauf, H. T. (2023). English language learning via YouTube: An NLP-based analysis of users' comments. *Computers*, *12*(2), 24. https://doi.org/10.3390/computers12020024

Anand, M., Sahay, K. B., Ahmed, M. A., Sultan, D., Chandan, R. R., & Singh, B. (2023). Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. *Theoretical Computer Science, 943*, 203-218. https://doi.org/10.1016/j.tcs.2022.06.020

Anjum, & Katarya, R. (2024). Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security, 23*(1), 577-608. https://doi.org/10.1007/s10207-023-00755-2

Arif, M., Shahiki Tash, M., Jamshidi, A., Ullah, F., Ameer, I., Kalita, J., ... & Balouchzahi, F. (2024). Analyzing hope speech from psycholinguistic and emotional perspectives. *Scientific Reports*, *14*(1), 23548. https://doi.org/10.1038/s41598-024-74630-y

Austin, D., Sanzgiri, A., Sankaran, K., Woodard, R., Lissack, A., & Seljan, S. (2020). Classifying sensitive content in online advertisements with deep learning. *International Journal of Data Science and Analytics, 10*(3), 265-276. https://doi.org/10.1007/s41060-020-00212-6

Balouchzahi, F., Sidorov, G., & Gelbukh, A. (2023). Polyhope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, *225*, 120078. https://doi.org/10.1016/j.eswa.2023.120078

Chakravarthi, B. R. (2022). Hope speech detection in YouTube comments. *Social Network Analysis and Mining*, *12*(1), 75. https://doi.org/10.1007/s13278-022-00901-z

Chakravarthi, B. R. (2022). Multilingual hope speech detection in English and Dravidian languages. *International Journal of Data Science and Analytics, 14*(4), 389-406. https://doi.org/10.1007/s41060-022-00341-0

Chinnappa, D. (2021). Dhivya-hope-detection@ LT-EDI-EACL2021: Multilingual hope speech detection for code-mixed and transliterated texts. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 73-78). Association for Computational Linguistics. https://aclanthology.org/2021.ltedi-1.11

Davidson, T., Bhattacharya, D., & Weber, I. (2019). *Racial bias in hate speech and abusive language detection datasets*. arXiv preprint arXiv:1905.12516.

Gowen, K., Deschaine, M., Gruttadara, D., & Markey, D. (2012). Young adults with mental health conditions and social networking websites: seeking tools to build community. *Psychiatric Rehabilitation Journal, 35*(3), 245. https://doi.org/10.2975/35.3.2012.245.250

Ghanghor, N., Ponnusamy, R., Kumaresan, P. K., Priyadharshini, R., Thavareesan, S., & Chakravarthi, B. R. (2021). IIITK@ LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil, Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 197-203). Association for Computational Linguistics.

Irfan, A., Azeem, D., Narejo, S., & Kumar, N. (2024). Multi-Modal Hate Speech Recognition Through Machine Learning. In *2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC)* (pp. 1-6). IEEE. https://doi.org/10.1109/KHI-HTC60760.2024.10482031

Kogilavani, S. V., Malliga, S., Jaiabinaya, K. R., Malini, M., & Kokila, M. M. (2023). Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings, 81*, 630-633. https://doi.org/10.1016/j.matpr.2021.04.102

Kumar, A. Saumya, S., & Roy, P. (2022). SOA_NLP@ LT-EDI-ACL2022: An ensemble model for hope speech detection from YouTube comments. In *Proceedings of the second workshop on language technology for equality, diversity and inclusion* (pp. 223-228). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.ltedi-1.31

Lee, Y., Yoon, S., & Jung, K. (2018). *Comparative studies of detecting abusive language on twitter*. arXiv preprint arXiv:1808.10245.

Louati, A., Louati, H., Albanyan, A., Lahyani, R., Kariri, E., & Alabduljabbar, A. (2024). Harnessing machine learning to unveil emotional responses to hateful content on social media. *Computers, 13*(5), 114. https://doi.org/10.3390/computers13050114

Malik, M. S. I., Nazarova, A., Jamjoom, M. M., & Ignatov, D. I. (2023). Multilingual hope speech detection: A Robust framework using transfer learning of fine-tuning RoBERTa model. *Journal of King Saud University-Computer and Information Sciences, 35*(8), 101736. https://doi.org/10.1016/j.jksuci.2023.101736

Mnassri, Kh., Farahbakhsh, R., Chalehchaleh, R., Rajapaksha, P., Jafari, A.R., Li, G., & Crespi, N. (2024). A survey on multi-lingual offensive language detection. *PeerJ. Computer Science*, *10*, e1934–e1934. https://doi.org/10.7717/peerj-cs.1934

Nagar, S., Barbhuiya, F. A., & Dey, K. (2023). Towards more robust hate speech detection: Using social context and user data. *Social Network Analysis and Mining, 13*(1), 47. https://doi.org/10.1007/s13278-023-01051-6

Nath, T., Singh, V. K., & Gupta, V. (2023). *BongHope: An annotated corpus for Bengali hope speech detection*. Research Square. https://doi.org/10.21203/rs.3.rs-2819284/v1

Palakodety, S., KhudaBukhsh, A. R., & Carbonell, J. G. (2020). Hope speech detection: A computational analysis of the voice of peace. In *ECAI 2020* (pp. 1881-1889). IOS Press.

RamakrishnaIyer LekshmiAmmal, H., Ravikiran, M., Nisha, G., Balamuralidhar, N., Madhusoodanan, A., Kumar Madasamy, A., & Chakravarthi, B. R. (2023). Overlapping word removal is all you need: Revisiting data imbalance in hope speech detection. *Journal of Experimental & Theoretical Artificial Intelligence*, *36*(8), 1837–1859. https://doi.org/10.1080/0952813X.2023.2166130

Roy, P., Bhawal, S., Kumar, A., & Chakravarthi, B. R. (2022, May). IIITSurat@ LT-EDI-ACL2022: Hope speech detection using machine learning. In *Proceedings of the Second Workshop on Language* Technology for Equality, Diversity and Inclusion (pp. 120-126). Association for Computational Linguistics. https://aclanthology.org/2022.ltedi-1.13

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1-10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101

Snyder, C. R., Rand, K. L., & Sigmon, D. R. (2002). Hope Theory: A Member of the Positive Psychology Family. In C. R. Snyder, & S. J. Lopez (Eds.), *Handbook of Positive Psychology* (pp. 257-276). Oxford University Press.

Subramanian, M., Sathiskumar, V. E., Deepalakshmi, G., Cho, J., & Manikandan, G. (2023). A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal, 80*, 110-121. https://doi.org/10.1016/j.aej.2023.08.038

Wang, Z., & Jurgens, D. (2018). It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 33-45). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1004

Yates, A., Cohan, A., & Goharian, N. (2017). *Depression and self-harm risk assessment in online forums*. arXiv preprint arXiv:1709.01848.

Yenala, H., Jhanwar, A., Chinnakotla, M. K., & Goyal, J. (2018). Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics, 6*, 273-286. https://doi.org/10.1007/s41060-017-0088-4

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