Sarcasm Detection

Алла Горбунова & Елизавета Клыкова

Актуальность проблемы

Голосовые ассистенты, чат-боты:

- классифицировать ввод пользователя, чтобы лучше понимать его намерения, настроение, реакцию
- классифицировать сгенерированные потенциальные ответы, чтобы не допустить двусмысленности, например, не похвалить пользователя фразой “*ну ты гений*”

Автоматическая модерация:

- например, токсичность может маскироваться сарказмом (“*вот она красотка конечно в этих леопардовых леггинсах*”)
- или наоборот: буквальное значение выглядит токсично, но его смысл противоположный (“*ну да, он же совсем дебил, сам не догадается закрыть окно*”)

Данные

Источник: SemEval 2022 (iSarcasmEval) – английские и арабские твиты, берем английские

Объем: 3468 твитов (867 саркастичных, 2601 не-саркастичных) + перифраз для саркастичных; включая перифраз, 3468 не-саркастичных, 4335 всего

Особенности: тексты размечены с точки зрения “сарказм / не сарказм” самими авторами, для каждого саркастичного текста приводится вариант без сарказма; есть fine-grained разметка саркастичных текстов: *sarcasm*, *irony*, *satire*, *understatement*, *overstatement*, *rhetorical question*

Достоинства: объективность разметки, т. к. разметчики – сами авторы текстов

Недостатки: предложения, перефразированные из сарказма, могут быть неестественными

Ссылка для скачивания

Задачи

Задача 1 и задача 3 из предложенных в рамках iSarcasmEval:

а) по тексту определить, саркастичный он или нет

> к этой задаче мы добавили подзадачу: проверить наше предположение о “неестественности” перефразированных несаркастичных текстов. Проведём два эксперимента: сначала в класс не саркастичных текстов включим только оригинальные твиты, затем возьмём их вместе с перифразами и сравним, с каким вариантом данных результат лучше.

б) внутри пары “сарказм + его перифраза” определить, какой из двух текстов саркастичный

Baseline

Классическое машинное обучение, признаки – tf-idf токенов. Мы много раз встречали такой подход в качестве базового и в статьях, и в учебных задачах. В подтверждение своих догадок нашли статью 2018 года, тоже SemEval, сходная задача – распознавание иронии: https://github.com/Cyvhee/SemEval2018-Task3/tree/master/benchmark_system

Там для классификации использовали LinearSVC, мы планируем провести эксперименты с разными классификаторами и использовать лучший результат как бейзлайн – для задачи а). Для задачи б) воспользуемся регрессией и будем присваивать лейбл по наибольшей вероятности, чтобы гарантировать ровно один положительный ответ в паре.

В качестве перестраховки будем держать в уме “рандомный” бейзлайн: соответствующий соотношению классов для задачи а) и 50/50 для задачи б).

Метрики

Для обеих задач считаем precision, recall, accuracy, F1-score. Главные метрики оставляем такие же, как в iSarcasmEval:

для а) – F1 для класса “сарказм”, потому что именно этот класс целевой и нам важна как точность, так и полнота, без перекосов;

для б) – accuracy, потому что распознавание происходит внутри пар, где всё сбалансировано по определению и accuracy = доля пар, где мы верно опознали сарказм.

План действий

- **Baseline:**
  - рандомный
  - для задачи а) – эксперименты с разными классификаторами, признаки – tf-idf токенов
  - для задачи б) – регрессия внутри пар и выбор максимального значения
- **Препроцессинг:**
  - лемматизация? lowercase? выбрасывать пунктуацию – точно плохая идея
  - различать ли “)” и “)))”, “!” и “!!!!” ? приводить ли случаи типа “lolllllll” к виду “lol”?
- **Основная модель:**
  - задача а) на корпусе без перифраз и с перифразами
  - задача б) для пар “сарказм – не сарказм”
  - предобученные эмбеддинги (с / без замораживания)
  - LSTM, CNN (на словах / символах), fine-tuned BERT – с упором на BERT, потому что LSTM и CNN делали в SemEval 2018, но интересно совместить и сравнить подходы
  - попробовать нагенерить фичи (есть ли капс, эмодзи)

Роли и задачи

Алла Горбунова

- ❏ идейный вдохновитель
- ❏ менеджер по трелло
- ❏ экспериментатор над бейзлайном
- ❏ задача б) (предварительно)

Елизавета Клыкова

- ❏ безыдейный исполнитель (шутка)
- ❏ оформитель презентаций
- ❏ экспериментатор над препроцессингом
- ❏ задача а) (предварительно)

Вместе

заполнение трелло, подготовка презентаций,
мозговой штурм основной части (задач а и б)

Литература

Основная идея: iSarcasmEval: Intended Sarcasm Detection In English and Arabic

Описание датасета для iSarcasmEval: Oprea, S., & Magdy, W. (2020). iSarcasm: A Dataset of Intended Sarcasm. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.118

Подтверждение, что предложенный бейзлайн адекватен для задачи такого типа: Van Hee, C., Lefever, E., Hoste, V. (2018). SemEval-2018 Task 3: Irony Detection in English Tweets. *Proceedings of The 12th International Workshop on Semantic Evaluation*. https://doi.org/10.18653/v1/S18-1005

Задача определения иронии на SemEval 2018 и подходы к ее решению: van Hee, C., Lefever, E., & Hoste, V. (2018). SemEval-2018 Task 3: Irony Detection in English Tweets. *Proceedings of The 12th International Workshop on Semantic Evaluation*. https://doi.org/10.18653/v1/s18-1005

Использование Densely Connected LSTM для задачи Irony Detection (SemEval 2018): Wu, C., Wu, F., Wu, S., Liu, J., Yuan, Z., & Huang, Y. (2018). THU_NGN at SemEval-2018 Task 3: Tweet Irony Detection with Densely connected LSTM and Multi-task Learning. *Proceedings of The 12th International Workshop on Semantic Evaluation*. https://doi.org/10.18653/v1/s18-1006

Важные ссылки

Github-репозиторий проекта

Доска в Trello (ссылка-приглашение)