**MIDTERM SUBMISSION**

Master in Intelligent Interactive Systems, Web Intelligence

Emma Fraxanet, David Moreno

1. **Project brief description**

The main idea is to retrieve movies or shows opinions from different sources (Twitter in the case of this midterm submission, Youtube, Reddit and Filmaffinity for the final submission) and analyze the reception of them before and after a movie release, or the evolution of a show along its chapters.

1. **Sentiment Analysis**

For sentiment analysis we follow the procedure:

* Clean text from stopwords + custom stopwords (such as https…)
* Study of TF & TF-IDF:
  + Useful in order to consider which are the most frequent words and how they could affect our wordclouds.
  + Useful to check the change in the different datasets, to see if these differences could be relevant.
* Use of VADER to extract polarity of tweets.
* First methods:
  + Separate tweets into positive and negative (exclude neutral tweets)
  + Treat positive and negative text separately:
    - Tokenize and remove words which appear in both lists (this words will be considered neutral in the coloring map)
    - Create a colored wordcloud of both texts combined, color map depending on the belonging of a word into a positive(green) or negative(red) tweet.The words that were in both texts appear colored in grey.
* Second method:
  + Use the full negative+positive tweets text.
  + The coloring will be done only by taking into account the polarity of the word.

*WANDAVISION PRELIMINAR ANALYSIS: SENTIMENT ANALYSIS*

From the TF-IDF analysis:

* We can see a difference throughout the episodes.
* The most used words make sense in the context of what happened in those episodes, the twitter users that are most active in commenting the show and other fan theories happening at the moment.
  + For example, in episode 5, where the ending had a big surprise revealed, it has a lot more mentions of the word “spoilers”.

From the colored Wordcloud. Regarding the two methods:

* First method: coloring regarding tweet polarity
  + The first method generally captures less negative words (especially in episode 3). We could think that maybe episode 3 was well received by the twitter community.
  + As we can see in episode 4, those tweets with negative polarity also use more curse words. We also see the name of another show that is releasing an episode every week, its color is red so the comparative might be negative. However, words such as “boring” are in grey and “magician” is in red, so the methodology should be polished.
  + For the 5th episode all words are from positive tweets or are neutral(in both text groups). There are no exclusively negative tweet words.
* Second method: coloring regarding individual word polarity
  + For the 3rd episode, we also have most words as neutral and positive. We feel it’s weird that spoilers are considered neutral. Marvel, which is the name of the producer company, is considered positive because of the meaning of the verb “to marvel”.
  + For episode 4, we again have curse words, along with more negative words. Boring is also considered a negative word. We think this episode might have had a not-so-good reception.
  + For episode 5 the negative words we fins (wtf, dead, bad, f\*\*k..) are probably related to the fact that the episode ended in a surprising way, and not that it was badly received (we think because in the previous method, all words were colored positive or neutral).

Observations: we could find a way to extract a well-received / bad received metric, we still don’t know how to do it exactly. We could also use LIWC.

*MALCOM&MARIE MOVIE BEFORE AND AFTER PREMIERE*

Not significant results (too general), probably due to:

1. Too few tweets
2. Not a complex enough sentiment analysis (only considering positive/negative)
   1. Possible alternative: LIWC
3. Tweets could be based on the description of the movie and not the personal opinion/review of the movie.
4. **Topic Modeling for premier releases**

For topic modelling what we try is to stream tweets with the keywords: ['premiere','new movie','new episode', 'new film'] every friday. From this we try to infer the different movies or episodes releases that are being talked about, and hopefully try to cluster the tweets into the different releases. For now, we have data of two days only, February 4th and February 5th, which we started using to test it.

The procedure follows: tokenizing, lemmatizing, creating bigrams, cleaning the data, eliminating links, usernames and stop words (usual plus custom such as ‘RT’..). We also do a study of which are the most frequent words, that would be in all tweets and therefore don’t help the clustering. Following that thought words that are less frequent also can be eliminated because they don’t help with the clustering, but disperse the topics too much. (Also, we saw that most “less frequent” words don’t even have a meaning). Finally we use the LDA Mallet Model, doing a coherence study to have a guide on how many topics should we build the model on.

**4febr:**

This day the premiere of a show called FastFoodies was released by truTV and they had a marketing add that asked people to tweet about the show in order to obtain a discount for a delivery food platform. This created a flooding of this tweets on the “premiere” stream. The coherence points at 14 topics for optimal number of topics, we ended up choosing 8 because it felt more ordered. However, most topics include this FastFoodies tweets (in red). The topic for the shows Wandavision and Riverdale , which are aired every thursday and friday, was also present (in blue), as well as some podcasts and other movies (in green).

**[(0,**

**[('tonight', 0.23235550392099913),**

**('hot', 0.20389195469067672),**

**('celebrate', 0.20360151031077547),**

**('meal', 0.20040662213186175),**

**('fastfoodie', 0.06622131861748475),**

**('omer', 0.0023235550392099913),**

**('performance', 0.0017426662794074934),**

**('cihangir', 0.0017426662794074934),**

**('burn', 0.0017426662794074934),**

**('chill', 0.0014522218995062445)]),**

**(1,**

**[('frame', 0.17370744010088274),**

**('analysis', 0.11191677175283733),**

**('video', 0.11034047919293821),**

**('show', 0.0980453972257251),**

**('propaganda', 0.09237074401008827),**

**('read', 0.09205548549810845),**

**('stop', 0.08890290037831021),**

**('chilling', 0.08669609079445145),**

**('game\_animation', 0.0028373266078184113),**

**('happy', 0.0025220680958385876)]),**

**(2,**

**[('fastfoodie', 0.49905541561712846),**

**('click', 0.36807304785894207),**

**('set', 0.003778337531486146),**

**('enjoy', 0.003778337531486146),**

**('cry', 0.0025188916876574307),**

**('role', 0.002204030226700252),**

**('bad', 0.002204030226700252),**

**('receive', 0.002204030226700252),**

**('break', 0.002204030226700252),**

**('unbelievable', 0.001889168765743073)]),**

**(3,**

**[('film', 0.06523096129837704),**

**('make', 0.0433832709113608),**

**('podcast', 0.031835205992509365),**

**('talk', 0.02902621722846442),**

**('listen', 0.020911360799001247),**

**('interview', 0.020911360799001247),**

**('live', 0.020599250936329586),**

**('today', 0.018414481897627965),**

**('time', 0.016853932584269662),**

**('join', 0.016229712858926344)]),**

**(4,**

**[('week', 0.10850923482849605),**

**('year', 0.06134564643799472),**

**('late', 0.05903693931398417),**

**('riverdale', 0.055408970976253295),**

**('wandavision', 0.03990765171503958),**

**('tomorrow', 0.034630606860158314),**

**('share', 0.03133245382585752),**

**('release', 0.030013192612137203),**

**('widely', 0.022427440633245383),**

**('brilliant', 0.021767810026385226)]),**

**(5,**

**[('hungry', 0.21596516690856313),**

**('chance', 0.20290275761973875),**

**('win', 0.14658925979680695),**

**('click', 0.10333817126269956),**

**('fastfoodie', 0.08679245283018867),**

**('doordash', 0.04586357039187228),**

**('gift', 0.02177068214804064),**

**('doorda', 0.020319303338171262),**

**('doord', 0.020029027576197386),**

**('doo', 0.01857764876632801)]),**

**(6,**

**[('movie', 0.14838709677419354),**

**('top', 0.1003225806451613),**

**('show', 0.1),**

**('attention', 0.09806451612903226),**

**('escape', 0.09806451612903226),**

**('expert', 0.09806451612903226),**

**('propaganda', 0.09580645161290323),**

**('starring', 0.008709677419354838),**

**('pm', 0.0064516129032258064),**

**('night', 0.005161290322580645)]),**

**(7,**

**[('good', 0.1931937172774869),**

**('free', 0.18193717277486912),**

**('true', 0.17984293193717277),**

**('sound', 0.17958115183246073),**

**('food', 0.17905759162303664),**

**('cli', 0.015445026178010472),**

**('yo', 0.013350785340314137),**

**('morning', 0.0015706806282722514),**

**('include', 0.0013089005235602095),**

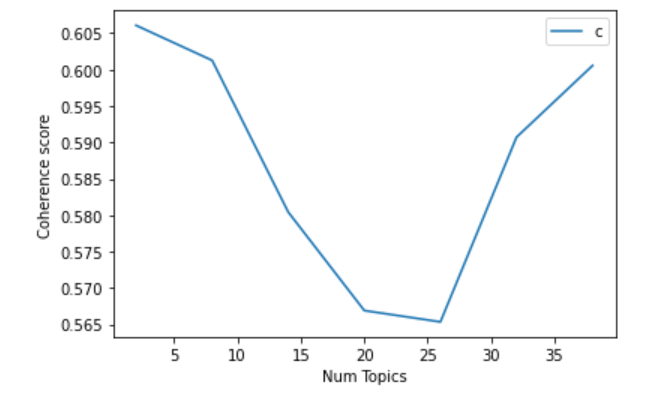
**('remake', 0.0007853403141361257)])]**

**Coherence Score: 0.5714980267461269**

**5febr:**

For this date the optimal number of topics is around 2. We can see that for this number of topics one topic is strongly about the show Wandavision, since this week there was a big surprise at the end and there was a lot of commenting about spoilers (also mentioned in the wordcloud part) and the other topic is about other releases in general (of movies, podcasts…). However one checking which documents are classified as each topic we see some considerable amount of misclassification still.

However, for around 8 topics we find a more disperse distribution (wandavision is mentioned in three different topics, but it is more accurate when detecting the other different topics (differences between podcasts and movies).

****

**Coherence Score: 0.6166562063083669**

**[(0,[('movie', 0.04728789986091794),**

**('film', 0.03529207232267038),**

**('watch', 0.02364394993045897),**

**('share', 0.018949930458970792),**

**('podcast', 0.01773296244784423),**

**('video', 0.012691237830319889),**

**('casting\_new', 0.012169680111265646),**

**('absolutely\_thrilled', 0.012169680111265646),**

**('talk', 0.011995827538247567),**

**('make', 0.011821974965229486)]),**

**(1,[('wandavision', 0.09158415841584158),**

**('stream', 0.04508486562942009),**

**('studio', 0.03978076379066478),**

**('week', 0.023514851485148515),**

**('spoiler', 0.021216407355021217),**

**('reaction', 0.018387553041018388),**

**('show', 0.014497878359264497),**

**('today', 0.012553041018387553),**

**('home', 0.011492220650636492),**

**('air', 0.010785007072135784)])]**

1. **Short term improvement**

As we have previously mentioned, some things are not working as well as expected, or not giving accurate results. Some short term improvements could be:

* Define a global metric for evaluating the sentiment of a whole episode (considering positive, neutral and negative tweets).
* Use LIWC (or similar) to get more precise sentiment from the tweets
* Test with more movies/shows

1. **Long term plans**

Regarding the future work of the project:

* Retrieve from YouTube the comments of movies/shows reviews from well-known review channels
* Retrieve Reddit opinions to movies/shows
* Retrieve Filmaffinity opinions to movies/shows
* Make the sentiment analysis for the same film from different resources and compare them to extract similarities and differences. Maybe we can conclude that some platforms are more optimistic/relaxed than others, maybe the topic modeling varies.