# Horoscope: tailored vagueness

Sofia Gervasoni

February 15, 2023

**Abstract**

Using different techniques of Text Mining & Sentiment Analysis, this paper aims to prove the generality of the horoscopes' texts. In the analysis of the sentiment for the different horoscopes, I applied two techniques: VADER and TextBlob. Thanks to this analysis, I proved the vagueness and neutrality of the language used by astrologists to write horoscopes. Moreover, it was possible to notice how the message delivered tends to be positive in most cases. Lastly, I used topic modelling, with which it was possible to confirm that the vocabulary used is always the same and common to all the horoscopes and zodiac signs.

## Contents

# 1 Introduction

How many times have you been stereotyped according to your zodiac sign? How often have you read your daily horoscope and thought: 'it is always right'? How often have you read the horoscope looking for a word of comfort? And you many times you found it? It probably happened so many times that you lost count. In this paper, I will show how it is possible.

This paper aims to understand the language used to write the horoscope and, in particular, analyse the text of the daily horoscope from Horoscope.com. Horoscope.com is the world's largest astrology media company with a portfolio of properties headlined by Astrology.com, Horoscope.com and the SunSigns mobile app.

## 1.1 Research questions

The hypothesis I wanted to prove and the research questions I wanted to answer are the following:

- According to Forer (a psychologist who proved his theory using personality tests), the horoscope would use general phrases. The aim is to impress a vast multitude of different people. The language used is neutral, vague and generic. Thanks to these characteristics, the description can be adapted to many people (under the guise of a personal, customised profile). Can we prove the truth of this fact? Do astrologers use generic language to write horoscope texts?
- According to MilanoPsicologo.it (Centro di Psicologia e Psicoterapia - Milan), the horoscope has a consoling effect. Horoscope has a good word for everyone and each zodiac sign is represented by desirable virtues. In short, it tickles our ego, our desire to be special, positive, and unique people. Among the words used to write the horoscope, can we find words with this function? In other words, is the polarity of the words mainly positive?
- The categories covered by the horoscope are more or less recurrent (i.e. work, love, life etc.), so for each daily horoscope, we expect to find words related to these topics. Will it be possible to distinguish these topics through the keywords used in each of them? Or will the vocabulary used always be the same and common to all, further confirming the generic nature of the horoscope?
- According to the various horoscope pages, however, the sun sign in the birth chart marks your personality and identity. It is the most powerful indicator of who you are in intimacy, in friendship, at work and in the various situations in your life. By studying the horoscopes of the different zodiac signs, is it possible to recognise predominant characteristics for each of them?

## 1.2 Scraping

I collected data through the Twitter API, a powerful tool that lets people scrape tweets directly on their local storage. First, I searched tweets through the username (@horoscopedotcom), taking all the tweets they posted from the 4th of February 2022 to the 4th of February 2023 (skipping the January 2023 ones). Unfortunately, the tweets were not complete, but they contained the URL to the website with the full daily horoscope for each zodiac sign. So, to obtain the full horoscope, I used the link to the horoscope's site. I collected 331 daily horoscopes for each zodiac sign, a total of 3.972 tweets (as the zodiac signs are 12: Aries, Taurus, Gemini, Cancer, Leo, Virgo, Libra, Scorpio, Sagittarius, Capricorn, Aquarius, and Pisces). In the end, I obtained a dataset with three columns: the daily horoscope, the sign and the date.

# 2 Pre-processing

The dataset contains 3.972 daily horoscopes (as one of these was a duplicate, I dropped it ), 269.771 words and 1.527.510 characters. The mean number of words used for each daily horoscope is 67.94, and the mean number of characters for each horoscope is 385.
First, I cleaned the horoscope texts by removing the non-letters and the words with only two characters. To avoid having problems with capitalization, I lowered all the words. After that, I used tokenization, stemming and lemmatization.

## 2.1 Tokenization

In this phase, I used the Natural language Toolkit (NLTK) tool. Tokenization means chopping up a defined document unit into pieces, called tokens. By doing so, I obtained a list of words for each horoscope text, where each word of the text becomes an element of the list. So I added to my dataset a column:'horoscope_token'.
At this point, I found out that the ten most common words are: 'could' (2581), 'may' (1923), 'might' (1767), 'day' (1239), 'time' (1214), 'take' (1179), 'get' (1104), 'make' (1038), 'people' (1035) and 'feel' (985). Looking at these very basic results it is possible to observe that the most common words used are related to the conditional form (i.e. 'could', 'may', 'might'). It seems that astrologists are not completely sure about what they are telling us. But let's keep on with the analysis and see if we can extract other interesting results.

## 2.2 Stop words

Stop words are a set of commonly used words in a language. Examples of stop words in English are 'a', 'the', 'is', 'are', 'and' etc. Stop words are commonly used in Text Mining and Natural Language Processing (NLP) to eliminate words that are so commonly used that they carry very little useful information. In this case, I noticed that each daily horoscope appoints the name of the zodiac sign to which it refers. So I decided to remove the name of the signs from all the horoscope texts. Moreover, since the horoscope is daily, the word 'today' was repeated many times, and since it is not significant to my analysis, I decided to drop it.

Then I normalized the tokens obtained converting each of them into their base form (base form = word - inflectional form). In doing so, I used two different techniques: lemmatization and stemming.

## 2.3 Normalization: Stemming and Lemmatization

Stemming is about chopping off the ends of words and this often includes the removal of derivational affixes. This indiscriminate cutting can be successful on some occasions, but not always. Another technique that could be used is lemmatization. Lemmatization is about removing inflectional endings of the words, returning the base or dictionary form of a word, which is known as the lemma. In doing so, the use of a vocabulary and morphological analysis of words are needed.
Both for stemming and lemmatization I used the NLTK tool, and in particular for stemming I used the Porter Stemmer. For the lemmatization, I also used the Spacy tool and I specified the part of sentences (POS) I wanted to keep for my analysis. In particular, I decided to keep only nouns and adjectives, whereas for topic modelling I compared the results obtained just by keeping nouns and adjectives, and the ones obtained using nouns, adjectives and verbs. At this point, I added two new columns 'tokens_no_stop_lem' (lemmatization with spacy tool) and 'tokens_no_stop_stem'.

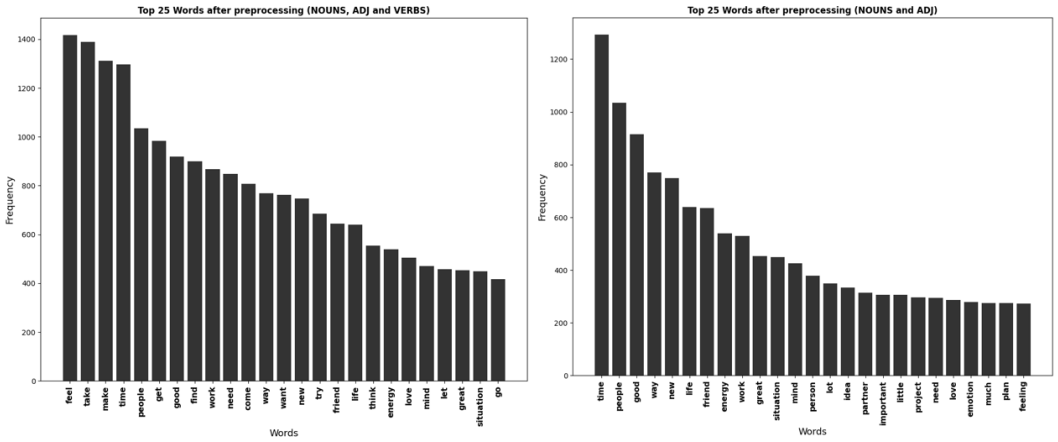


Figure 1: Most common words after pre-procesing

# 3 Sentiment analysis

Answering the first and second research questions, I implemented sentiment analysis. From this analysis, I expect to find a slightly positive sentiment. According to the hypothesis expressed above, the horoscope uses general and neutral words, however, it tends to deliver a positive message, never being too extreme, either in a positive or negative sense. To implement sentiment analysis, I decided to compare the results obtained with two different methods: VADER and TextBlob.

## 3.1 VADER

With VADER, I obtained four results: the percentage of positive words in each horoscope (pos), the percentage of negative words in each horoscope (neg), the percentage of neutral words in each horoscope (neu) and the compound score. The compound score is the sum of positive, negative neutral scores, which is then normalized between -1 (most extreme negative) and +1 (most extreme positive). When the compound score is close to +1 the sentence expresses an extremely positive sentiment.

I plotted the compound score with a barplot (Figure. 2)and I noticed that the majority of the scores are very close to 1 (extremely positive). Nevertheless, analysing these results and reading the Horoscopes, which VADER classifies as extremely positive, I noticed that these sentences were not as positive as VADER says. For example, the Horoscope with the highest positive compound score (0.9929) for VADER is:

*"Remember that the important thing isn't necessarily what you're doing but the people you're with, Leo. There's a great deal of passion in the air today that you can latch onto and put to good use. Have fun and remember to smile. You can make a great deal of progress toward your goals as long as you stay motivated. Connect with others and feel the strength of shared resources".*

In my opinion, this sentence for sure doesn't express a negative sentiment but on the other hand, it doesn't express a very positive sentiment. So, to me, it seems that this sentence is slightly positive, not extremely positive. In this sentence, 43.6% of words are positive, 56.4% are neutral and 0.00% are negative. So the particularly positive compound score is due to the fact that in this sentence there are no negative terms.

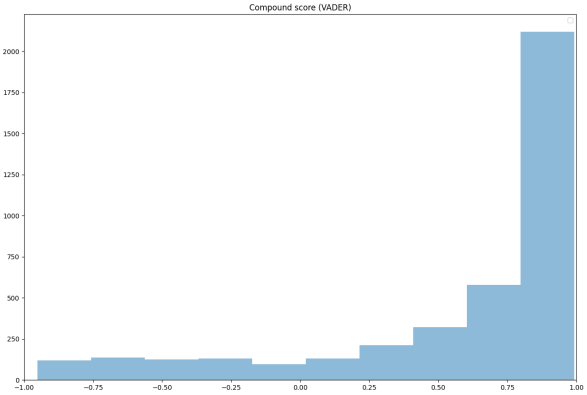


Figure 2: Compound score - VADER

I do not agree with these results as, in my opinion, they appear too optimistic. The mean of the percentage of positive words among all the horoscopes is 17.4%, for the negative the mean percentage is 5.6%, whereas for the neutral it is 77%. So even if the percentage of positive words is higher than the percentage of negative words (in mean), the percentage of neutral words is much higher.

## 3.2 TextBlob

As I was not satisfied with the results obtained with VADER, I decided to try with TextBlob. TextBlob goes along finding words and phrases it can assign polarity and subjectivity to, and it averages them

all together for longer text. As VADER, TextBlob is a lexicon method (the lexicon approach has a mapping between words and sentiment, and the sentiment of a sentence is the aggregation of the sentiment of each term), but TextBlob seems to work better with formal texts.
As for VADER, I plotted the results in a barplot (Figure.3), and in this case, most of the results range between -0.5 and 0.75, with a peak in 0.25. According to this method, in general, the sentiment is neutral, or better, slightly positive. In my opinion, these results seem much more realistic than the one obtained with the previous method and totally according to the hypothesis I wanted to prove.

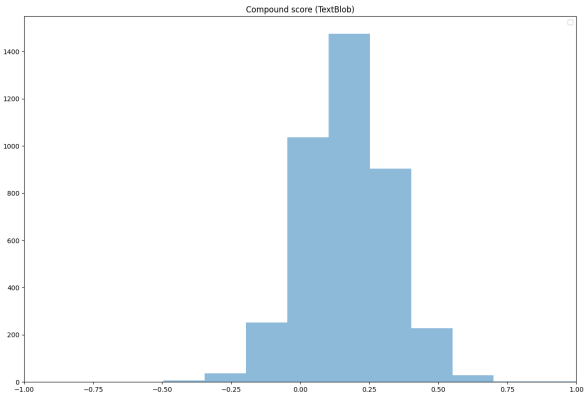


Figure 3: Compound score - TextBlob

## 3.3 VADER vs. TextBlob

To compare the two methods, I decided to plot the compound scores obtained with the two methods in a scatter plot. On the x-axis, I plotted the results obtained with TextBlob, whereas on the y-axis I plotted the results obtained with VADER. As you can see from the plot (Figure.4), there is a discordance between the results obtained with the two methods, as many scores are located in the 2nd and 4th quadrants (meaning that when the compound score is positive for TextBlob is negative for VADER and vice-versa). As seen previously, VADER seems to emphasise too much the positive results, whereas TextBlob returns more reasonable results.

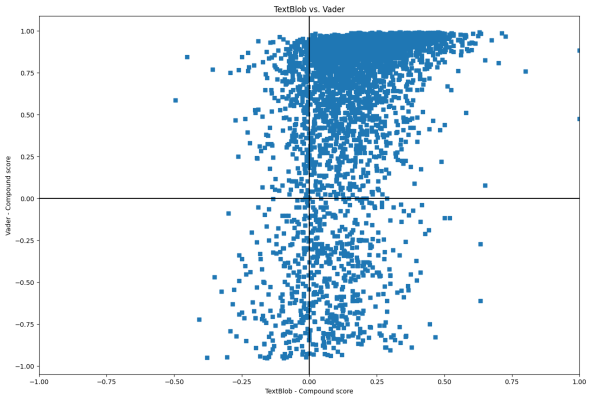


Figure 4: Vader VS TextBlob

These results prove the hypothesis expressed in the first and second research questions. The tone used to write the horoscope is neutral and tends to be neutral, vague and generic language. Nevertheless, most of the time delivers a positive message to the reader, and even if generic, it tends to be positive rather than negative. This language makes the reader feel unique and special, delivering positive vibes. Even when the horoscope is negative, it tends to be neutral rather than negative.

# 4 Topic modelling

To answer the third research question (see paragraph 1.1), identify the main topics treated by horoscope and recognize the most common words used in all these topics, I decided to use the topic modelling technique. This technique is used to identify which topic is discussed in a document. With this purpose, I applied Latent Dirichlet Allocation (LDA), which is used to convert a set of sentences into a set of topics. This unsupervised (no response variable) machine-learning method helps us discover hidden semantic structures in a document. To carry out the analysis, I used the Gensim tool.
We define a topic as a collection of dominant keywords that are typical representatives. Just by looking at the keywords, we identify what the topic is all about.

First of all I created bigrams and trigrams, to see if there are any words that recurrently repeat combined. In particular, bigrams are two words frequently occurring together in the document and trigrams are three words frequently occurring. To make easier the reading of these words, I plotted them in Figure 5. In this graph, it is possible to see how words like 'romance', 'love', 'romantic' and 'partner' are used together in the horoscopes, and this could be a topic (relationships). Other topics could be related to 'self-love', 'hard-work, 'family members' and others.

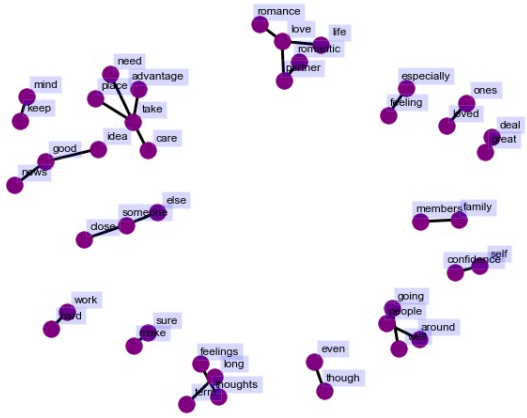


Figure 5: N-grams

## 4.1 Nouns and adjectives

I run the first analysis taking into consideration just nouns and adjectives (dropping verbs and adverbs). In the following subsection, I will show the results obtained keeping also the verbs.
I created the dictionary and the corpus which are two elements needed to run topic modelling. In particular, the corpus tells us the term document frequency and takes the form (word_id, word_frequency). For example, (2, 1) implies that the word id 2 occurs once in the document we are analysing.
At this point it is necessary to identify the optimal number of topics that should be kept in the model. For this purpose, I calculated the coherence values obtained with a different number of topics (k) and pick the one that gives the highest coherence value. Coherence scores a single topic by measuring the degree of semantic similarity between high-scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artefacts of statistical inference.
For computational reasons I decided to give k values ranging **between 2 and 20**, with **step=2**, and I plotted the results in Figure 6. According to this plot, the ideal number of topics is 12 (Coherence Value equal to 0.3873).
Choosing this number of topics seems not that significant when it comes to the analysis as 98% of the horoscope is collected in one topic and the other 2% split between the other 11 topics. In Figure 7, I introduced a table containing the percentage of horoscopes for each topic and the keywords related to these latter. As we can see from this table, the majority of the horoscopes fall in topic 3. Looking at

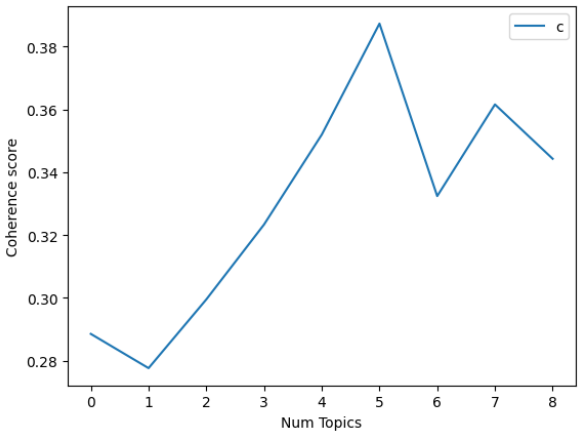


Figure 6: Coherence Score by number of topics

the keywords for this topic, my attention falls in particular on three of these: 'life', 'friend' and 'work'. These categories (together with 'love') are usually the main categories we read in the horoscopes. And what about 'love'? Isn't it one of the main topics treated by horoscopes? Well, I would say that in this case 'love' is treated separately from the other three categories listed above. Actually, if we look at the second most common topics in our general daily horoscopes, it is possible to notice the presence of four words in particular: 'partner', 'feeling', 'relationship' and 'romantic'. These latter words are strongly related to love (as it is possible to notice from Figure 5), so it is possible to say that also love is treated in our daily horoscopes but it is treated separately from the other topics. Meaning that if a daily horoscope speaks about love, it would not treat other topics like 'work' or 'friends' but it would focus on 'love'.

| Topic | Percentage | Topic_Keywords |
|---|---|---|
| 3 | 0.9874 | time, good, way, new, life, friend, energy, work, situation, person |
| 5 | 0.0086 | people, great, important, plan, partner, feeling, relationship, romantic, day, usual |
| 1 | 0.0035 | thing, other, emotion, strong, action, able, free, fact, ability, use |
| 8 | 0.0003 | bit, close, issue, problem, difficult, head, clear, try, hard, point |
| 9 | 0.0003 | sure, dream, emotional, attention, careful, truth, question, talk, busy, honest |

Figure 7: Dominant topics KeyWords

## 4.2 Nouns, adjectives and verbs

I proceeded in the same way as in the previous analysis and, in this case, the ideal number of topics seems to be four. According to Figure 8, the peak for the coherence score is obtained with k=4, where the coherence score is 0.3837.

Also in this case I resumed the keywords for the dominant topics in a table (Figure 9). In this case, the dominant topic is Topic 1, which is present in almost 77% of the horoscopes. In my opinion, this topic is generic as it is difficult to extract a topic from these keywords. The second dominant topic (Topic 3) is present in 11% of the horoscopes. From the keywords for this topic, it is easier to extract a topic. Actually, we have words like 'love', 'feeling' and 'partner' which suggest as topic romantic relationships/love. The third dominant topic, in this case, is Topic 0, which appears in 7% of the horoscopes. This topic seems mostly related to what will happen in the future in terms of friendships and interests. The last dominant topic is Topic 2, which appears in 4% of the horoscopes. Even in this case, it is hard to find the topic connected to the keywords as they are generic and hard to combine. Using the verbs seems that the results become even more generic and hard to define.

From these results (both including verbs or not) it is possible to further confirm the generality of the horoscopes. It is actually possible to notice the difficulty to distinguish the horoscopes between

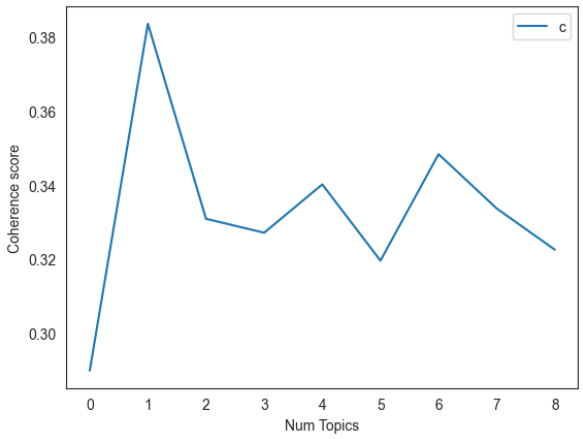


Figure 8: Coherence Score by number of topics

| Dominant_Topic | Percentage | Topic_Keywords |
|---|---|---|
| 1 | 0.7668 | get, take, make, time, go, need, good, work, way, find |
| 3 | 0.1174 | re, feel, life, love, person, look, great, feeling, partner, people |
| 0 | 0.0728 | new, friend, lot, idea, expect, involve, come, future, want, interest |
| 2 | 0.0431 | people, other, situation, try, emotion, thing, issue, fact, word |

Figure 9: Dominant topics KeyWords

the topics as most of the horoscopes fall on the same topic. Answering the research question above, from these results we can confirm that the vocabulary used is always the same and common to all the horoscopes, further confirming the generic nature of the horoscope.

# 5 Most common adjectives for each sign

To answer the last research question and see if it is possible to infer some characteristics of the zodiac signs in the daily horoscopes analysed, I decided to keep only adjectives for this analysis. The following table contains the most common adjectives for each sign.

| **Zodiac sign** | **Most common adjectives** |
|---|---|
| Aquarius | romantic, strong, creative, physical, social, emotional |
| Aries | strong, artistic, romantic, creative, positive, free, social |
| Taurus | strong, romantic, clear, careful, emotional |
| Gemini | romantic, strong, interesting, free, positive, careful |
| Cancer | careful, emotional, strong, positive, creative |
| Leo | strong, social, creative, romantic, spiritual, emotional |
| Virgo | creative, spiritual, artistic, positive, romantic, sensitive, physical, strong |
| Libra | romantic, strong, social, creative, careful, passionate |
| Scorpio | strong, romantic, positive, emotional, creative |
| Saggitarius | romantic, free, careful, spiritual, strong, creative |
| Capricorn | romantic, strong, sensitive, physical, free, social, careful |
| Piscis | strong, free, emotional, creative, powerful, romantic |

Looking at the most common adjectives used in the daily horoscopes for the different signs, it is possible to notice that most of them are repeated for all the signs. This is a further confirmation of the randomness and generality of the horoscopes.

# 6 Conclusions

Using Text Mining and Sentiment Analysis, I answered all the research questions proposed in section 1.1.

Thanks to sentiment analysis, I proved the vagueness of the language used to write the horoscopes. According to the theory proposed by Forer, astrologists use general sentences, aiming to impress a vast multitude of different people. The language used is neutral, vague and generic. Thanks to these characteristics, the description can be adapted to many people. Moreover, as sustained by MilanoPsicologo.it (Centro di Psicologia e Psicoterapia - Milan), the horoscope has a consoling effect, so it always tends to deliver a positive message.

Using topic modelling I was able to further confirm the generality of the horoscopes, addressing the third research question. It results hard to distinguish the horoscopes between the topics, as most of the horoscopes fall on the same topic. So, from the results obtained with topic modelling, I can confirm that the vocabulary used is always the same and common to all the horoscopes.

Lastly, collecting the most common adjectives used to write the daily horoscopes for the different signs, I noticed that most of these adjectives are used for all the zodiac signs, making it impossible to recognise the predominant characteristics of each one.